ANNALS OF MATHEMATICS
Vol. 40, No. 1, January, 1939

NOTES ON LINEAR TRANSFORMATIONS. II. ANALYTICITY OF SEMI-GROUPS[1]

BY EINAR HILLE

(Received June 3, 1938)

1. PROBLEM AND METHODS

1.1 Introduction

In the first note of this series (see E. Hille [2]) the author showed that certain well-known functional transformations of the form

$$U_\lambda x = \lambda \int_{-\infty}^{\infty} K(\lambda u) x(t+u)\, du \tag{1.1.1}$$

satisfy functional equations with respect to the parameter λ.

In particular, for suitably restricted classes of functions $x(t)$, the two transformations[2]

$$W_\alpha x = (\pi\alpha)^{-\frac{1}{2}} \int_{-\infty}^{\infty} \exp(-u^2/\alpha) x(t+u)\, du, \tag{1.1.2}$$

and

$$P_\alpha x = \frac{\alpha}{\pi} \int_{-\infty}^{\infty} \frac{x(t+u)}{u^2+\alpha^2}\, du, \tag{1.1.3}$$

were found to satisfy the same equation

$$T_\alpha[T_\beta x] = T_{\alpha+\beta} x, \qquad \alpha, \beta > 0. \tag{1.1.4}$$

Here $W_\alpha x$ is the Gauss-Weierstrass singular integral and $P_\alpha x$ is the Poisson integral for the upper half-plane. Poisson's integral for the circle leads to the same functional equation provided we replace the customary parameter r by $e^{-\alpha}$. The equation is also encountered in the applications of Abel's method of summation or its various generalizations (convergence factors $e^{-\alpha\lambda_n}$). As an example of a different type, leading to the same equation, we mention the theory of fractional integration.

It was observed in my paper [2] that $W_\alpha x$ and $P_\alpha x$ are analytic functions of the parameter in the right half-plane and the same property holds in the case

[1] Presented to the American Mathematical Society, February 26 and April 16, 1938. A preliminary communication (see the list of References at the end of this paper, E. Hille [3]) was made to the National Academy of Sciences on February 9, 1938. The results of this note are developed more fully in §§4.1 and 4.2 of the present paper.

[2] These transformations can evidently be reduced to the form of (1.1.1) by a suitable change of the parameter.

of the other examples just mentioned. These instances are sufficiently important to warrant an investigation of the analyticity properties of the solutions of (1.1.4). A first attack upon this problem will be made in the present paper.

1.2. Problems and Results

We consider an abstract metric space E of Banach's type and a one-parameter family of bounded transformations T_α on E to E. The parameter α is restricted to a real or complex point set Z. *If T_α satisfies* (1.1.4) *for all α, β, and $\alpha + \beta$ in Z, we say that T_α has the semi-group property in Z. In particular, we call the family a semi-group if its domain of definition Z is invariant under addition and $\alpha = 0$ is a non-isolated point of the domain where $T_0 = I$.*

Our problem is to determine sufficient conditions under which the family has the semi-group property in a domain Z and T_α is a holomorphic function of α in Z in a sense to be made precise later. In the present paper we concentrate the attention on one feature of this problem. We suppose that a semi-group is defined on the positive real axis and ask when it admits of an analytic extension to the complex plane preserving the semi-group property. As is to be expected, the last proviso is superflous; an analytic extension always preserves the semi-group property (Theorem 5 below). We limit our problem still further by being interested only in the case in which an analytic extension exists in a right half-plane. This case seems to be the most important one and is fairly easy to characterize in simple terms.

It is not difficult to give examples of semi-groups which are not measurable, bounded or continuous with respect to the parameter. In a study of the analyticity problem it is necessary to exclude these pathological cases. Even if this is done, analytic extensions do not have to exist. It appears from the results of this paper that a decisive rôle is played in the problem by the properties of the resolvent $R(\lambda, \alpha) = [\lambda I - T_\alpha]^{-1}$ for small positive values of α.

Since T_α is a bounded transformation by assumption, the spectrum $S(\alpha)$ of T_α is restricted to a neighborhood of the origin in the complex λ-plane. If for small values of α the spectrum of T_α avoids completely a fixed sector $\psi < \arg \lambda < 2\pi - \psi$, $0 < \psi < \pi/2$, and as $\lambda \to 0$ in this sector the norm of $R(\lambda, \alpha)$ does not grow faster than a negative power of $|\lambda|$, the exponent of which may vary with α but is at most $O(1/\alpha)$, then there exists an analytic extension of the semi-group in a half-plane. If the exponent is $o(1/\alpha)$ instead, this half-plane is the right half-plane and the extension is also a semi-group.

As a rule we are only interested in analytical semi-groups, but we note that if for small α's the spectrum of T_α can be enclosed in a fixed circle which leaves $\lambda = 0$ on the outside, then the original semi-group can be extended to the whole plane and generates a group.[2a]

[2a] [Added in proof:] Cf. also the author's paper "Analytical semi-groups in the theory of linear transformations" in the forthcoming Proceedings of the Ninth Scandinavian Mathematical Congress.

The paper is divided into four chapters. The first is devoted to the general theory of one-parameter families of transformations. The first two paragraphs contain a discussion of the underlying topologies and the analysis of functional operators. There is probably not much which is strictly new here, but the results are needed for the later discussion and they are included for the convenience of the reader. We are mainly interested in families analytic in the parameter and of exponential order of growth in a right half-plane. Such transformations are representable by various classes of operational binomial series the theory of which is outlined in §2.3. Families with the semi-group property appear in the last two paragraphs of the chapter. In §2.4 I have collected all material known to me concerning the relationships between the semi-group property on one hand and measurability, boundedness, and continuity on the other (see also §5.2). In §2.5 analyticity is brought in.

Chapter 3 is devoted to analytical semi-groups bounded in a half-plane. Here we use the term bounded in the sense of exponential order of growth of the norm. In §3.1 we show that holomorphism and boundedness in a half-plane of a semi-group impose severe restrictions on the resolvent $R(\lambda, \alpha)$ and the spectrum $S(\alpha)$ of T_α for small positive values of α. A further study of the spectral properties of the transformations in a semi-group is given in §§3.2 and 3.4.

The main theorem, already sketched above, is proved in §3.3. We show that if a semi-group $\{T_\alpha\}$, defined for $\alpha > 0$, and satisfying mild restrictions as to measurability of the norm and continuity of the transformations, in addition, has the property that the resolvents and spectra are limited in the manner indicated above for small values of α, then there exists an analytic extension in a right half-plane. For this purpose we study the integral

$$W(\zeta, \alpha) = \frac{1}{2\pi i} \int_C R(\lambda, \alpha)\lambda^{\zeta/\alpha} d\lambda,$$

where the contour of integration surrounds the spectrum of T_α, beginning and ending at the origin. The integral defines an operator function holomorphic in a right half-plane which for $\zeta = n\alpha$ equals $T_{n\alpha}$. We can always form this integral which interpolates the sequence of operators formed by the integral powers of T_α. Thus the crux of the problem is to show that $W(\zeta, \alpha)$ is independent of α and agrees with T_ζ for sufficiently large real values of ζ. This succeeds with the aid of an important theorem due to F. Carlson [1], see §5.1. The theory of binomial series plays a conspicuous rôle in this chapter.

The case in which the spectrum of T_α, $\alpha > 0$, is always positive and the resolvent is of finite order for approach of λ towards the spectrum is particularly interesting (§3.5). Here we have analytic extensions representable by various classes of generalized Laplace-Stieltjes integrals. Finally we prove an embedding theorem in §3.6: any bounded transformation whose spectrum is enclosed in a sector of opening less than π and vertex at the origin and whose

resolvent is of finite order outside of this sector can be embedded in a family having the semi-group property. This family is formed by the powers of T.

Chapter 4 gives various special instances of the general theory: self-adjoint transformations in Hilbert space, the Gauss-Weierstrass and Poisson transformations, Abel summability, and fractional integration.

During the preparation of this manuscript I have profited much from discussions with Professors N. Dunford and J. D. Tamarkin whose constructive criticism has contributed much to the paper in particular in §2.2.

2. One-Parameter Families of Transformations

2.1. The Basic Topologies

We shall be concerned with transformations defined in an abstract space E. We assume that E is a space (B) in the terminology of Banach, i.e., E is a normed linear vector space, complete with respect to its metric. The elements of E will be denoted by x, y, etc. The norm of x is denoted by $\| x \|$. All complex numbers are admitted as multipliers of the elements of E.

Together with the space E we also consider the conjugate space $\bar{E}$ whose elements are the bounded linear functionals L on E. We recall that $\bar{E}$ can be metrized by taking $\| L \| = \sup_{\|x\|=1} | L(x) |$.

The metric defines the strong topology in E. We have also the weak topology in which the relation $x_n \to x_0$ means that $L(x_n - x_0) \to 0$ for every $L \in \bar{E}$.

The set of all linear bounded transformations U on E to E forms a ring $\mathfrak{S}$, non-commutative and having divisors of zero. We recall that there are three topologies of special interest in $\mathfrak{S}$, viz. the weak, the strong, and the uniform topologies.[3] In the weak topology the relation $U_n \to U_0$ means that $L((U_n - U_0)x) \to 0$ for each $x \in E, L \in \bar{E}$. In the strong topology it means $\| (U_n - U_0)x \| \to 0$ for each $x \in E$, and in the uniform topology $\| U_n - U_0 \| \to 0$, i.e., $\sup_{\|x\|=1} \| (U_n - U_0)x \| \to 0$.

2.2. The Analysis of Functional Operators

In our study of semi-groups we shall have to apply various processes of analysis to families of linear transformations in their dependence upon parameters. For the convenience of the reader we collect in this and the next paragraphs what we shall need in the way of notions and results.

Let $U(\zeta)$, $\zeta = \xi + i\eta$, denote a family of transformations in $\mathfrak{S}$, defined for ζ in a set Z of the complex ζ-plane. We say that $U(\zeta)$ is *weakly measurable* in Z, if $L(U(\zeta)x)$ is a measurable function of ζ in Z for each $x \in E$, $L \in \bar{E}$. We say that $U(\zeta)$ is *continuous* at $\zeta = \zeta_0$ in the sense of one of the topologies (weak, strong or uniform) if the relation $U(\zeta_n) \to U(\zeta_0)$ holds in the sense of that topology whenever $\zeta_n \to \zeta_0$. The meaning of continuity in a domain and in the sense of the various topologies is obvious. We note that uniform continuity implies strong continuity which in turn implies the weak kind. If Z is a set

[3] These notions are due to J. von Neumann [1].

on the real axis we can also define continuity on the right or on the left, semi-continuity etc. Thus, in particular, $U(\xi)$ is strongly continuous on the right at $\xi = \xi_0$ if $\| [U(\xi_0 + h) - U(\xi_0)]x \| \to 0$ for each $x \in E$ as $h \to 0$ through positive values.

Again if $U(\xi)$ is defined for $\alpha \leqq \xi \leqq \beta$ we can define the notions of bounded variation. Thus, for example, $U(\xi)$ is of bounded variation in the strong sense in $[\alpha, \beta]$ if for any sub-division of $[\alpha, \beta]$ and any $x \in E$

$$\sum_{k=1}^{n} \| [U(\xi_k) - U(\xi_{k-1})]x \| \leqq M(x).$$

We shall need several notions of integration. The integrals will either be contour integrals in the complex plane or Riemann-Stieltjes integrals on an interval (α, β). The integrand will be either an operator function $U(\zeta)$ in $\mathfrak{S}$ or elements $U(\zeta)x$ in E or bilinear functionals $L(U(\zeta)x)$ in the product space $(E, \bar{E})$.

In the case of an operator integral we consider sums of the usual type

$$U_n = \sum_{k=1}^{n} U(\zeta_{k,n}^{*})(\zeta_{k,n} - \zeta_{k-1,n}).$$

Here the ζ's are points on a rectifiable oriented arc C in the complex plane and $U(\zeta)$ is defined on C. Suppose that as $\max | \zeta_{k,n} - \zeta_{k-1,n} | \to 0$ the sums converge (in the weak, strong or uniform sense) to a limit independent of the mode of sub-division. This means that there exists an operator U^* in $\mathfrak{S}$ such that U_n converges to U^* in the topology in question. We refer to U^* as the *weak, strong or uniform integral*, as the case may be, of $U(\zeta)$ along C and use the notation $\int_C U(\zeta)d\zeta$, leaving the character of the integral to be surmised from the context. The space $\mathfrak{S}$ is complete in the strong as well as in the uniform sense, but may conceivably not be weakly complete. This fact may sometimes render the notion of weak convergence somewhat illusory.

Thus if $U(\zeta)$ is strongly continuous along C, the sums U_n converge strongly, i.e., $U(\zeta)$ is integrable in the strong sense along C. The same is true if we replace the word "strong" throughout by the word "uniform." But if we know that $U(\zeta)$ is weakly continuous along C, we can merely show that $L(U_n x)$ converges to a bilinear functional $F(x, L)$ on the product space $(E, \bar{E})$, and in general we do not know for sure that $F(x, L)$ is of the form $L(U^*x)$ with $U^* \in \mathfrak{S}$. This complicates matters somewhat, but, fortunately, in the case of greatest interest to us, that in which $U(\zeta)$ is holomorphic, the three notions of integration coincide as we shall see below.

The strong convergence of the Riemann sums formed out of the operators $U(\zeta)$ on C in $\mathfrak{S}$ implies and is implied by the strong convergence of the Riemann sums formed out of the corresponding elements $U(\zeta)x$ of E. Thus we have also integrals of the type $\int_C [U(\zeta)x]d\zeta$ and

$$\left[\int_C U(\zeta)d\zeta\right]x = \int_C [U(\zeta)x]d\zeta. \tag{2.2.1}$$

Here both integrals are taken in the strong sense, but if the integral on the left exists in the weak sense, so does the integral on the right. The relation

$$(2.2.2) \qquad L\left\{\left[\int_c U(\zeta)\,d\zeta\right]x\right\} = L\left\{\int_c [U(\zeta)x]\,d\zeta\right\} = \int_c L[U(\zeta)x]\,d\zeta$$

holds for weak integrals. For strong integrals we have further

$$(2.2.3) \qquad \left\| \int_c [U(\zeta)x]\,d\zeta \right\| \leqq \int_c \| [U(\zeta)x] \| \, | \, d\zeta \, | \, ,$$

and for uniform integrals even

$$(2.2.4) \qquad \left\| \int_c U(\zeta)\,d\zeta \right\| \leqq \int_c \| U(\zeta) \| \, | \, d\zeta \, | \, .$$

We leave it to the reader to develop the theory of Riemann-Stieltjes integrals. We shall merely be concerned with the case in which the integrand is uniformly continuous, the integrator is uniformly of bounded variation and the limit of the Riemann-Stieltjes sum is taken in the uniform sense.

We say that $U(\zeta)$ is *differentiable at* $\zeta = \zeta_0$ in the sense of the topology if $\lim_{h\to 0}[U(\zeta_0 + h) - U(\zeta_0)]/h$ exists in the sense of that topology and is independent of the mode of approach to zero of h. If $U(\zeta)$ is uniquely defined, continuous, and differentiable at all points of a domain Z, we say that $U(\zeta)$ is *holomorphic* in Z in the sense of the topology in question. It is clear that uniform holomorphism implies the strong kind which in turn implies the weak one. But here the converse is also true and even still more.

THEOREM 1. *If an operator function $U(\zeta)$ has the property that $L[U(\zeta)x]$ is holomorphic in a fixed domain Z for every $x \in E$, $L \in \bar{E}$, then $U(\zeta)$ is holomorphic in Z in the uniform sense.*[4]

PROOF. Let Z_0 be any bounded domain interior to Z and having a positive distance from the complementary set of Z. We shall base our proof upon the fact that if a numerical function is holomorphic in Z its difference quotient $[f(\zeta + h) - f(\zeta)]/h$ converges uniformly in Z_0. Thus we can find a constant $M = M(x, L)$ such that for ζ, $\zeta + h$, and $\zeta + g$ in Z_0

$$| \, L\{([U(\zeta + g) - U(\zeta)]/g - [U(\zeta + h) - U(\zeta)]/h)x\} \, | \leqq | \, g - h \, | \, M(x, L).$$

But a theorem of Banach ([1], p. 80, theorem 6) says that if the least upper bound of $| \, L(y) \, |$ for y in a fixed set $Y \in E$ is finite for every L in $\bar{E}$, then Y must be a bounded set. Letting Y be the set of elements

$$([U(\zeta + g) - U(\zeta)]/g - [U(\zeta + h) - U(\zeta)]/h)x/(g - h),$$

where x is fixed and ζ, $\zeta + g$ and $\zeta + h$ range over Z_0, we see that we can find a constant $N = N(x)$ such that

$$\| \, ([U(\zeta + g) - U(\zeta)]/g - [U(\zeta + h) - U(\zeta)]/h)x \, \| \leqq | \, g - h \, | \, N(x).$$

[4] This theorem is due to Professor Nelson Dunford. I publish it here with his permission.

This inequality shows that the difference quotient $[U(\zeta + h) - U(\zeta)]/h$ converges strongly in Z_0. Since $\mathfrak{E}$ is complete in the strong sense we conclude that $U(\zeta)$ is strongly differentiable in Z_0 and consequently also holomorphic in the strong sense. Z_0 being arbitrary, this conclusion extends to Z.

But we can go further. Another theorem of Banach ([1] p. 80, theorem 5) says that if the least upper bound of $\| Vx \|$ for V ranging over a fixed set $\mathfrak{E}_1 \subset \mathfrak{E}$ is finite for every fixed x in E, then $\mathfrak{E}_1$ must be bounded in the uniform topology of $\mathfrak{E}$. This proves the existence of a constant C such that

$$\| [U(\zeta + g) - U(\zeta)]/g - [U(\zeta + h) - U(\zeta)]/h \| \leqq C \, | g - h \, |$$

for ζ, $\zeta + g$, and $\zeta + h$ ranging over Z_0. From this we conclude uniform differentiability in Z_0 and consequently that $U(\zeta)$ is holomorphic in Z in the uniform sense. This completes the proof of the theorem. Thus the three notions of holomorphism coincide, and we can say simply that $U(\zeta)$ is holomorphic in Z without specifying the topology.

This implies that if C is a simple rectifiable arc in Z and $U(\zeta)$ is holomorphic in Z, then the integrals $\int_C U(\zeta)d\zeta$ and $\int_C [U(\zeta)x]\, d\zeta$ exist as uniform limit in $\mathfrak{E}$ and strong limit in E respectively. But the theorem also gives us a convenient means of extending ordinary complex function theory to operational function theory in the strong topology of E and the uniform topology of $\mathfrak{E}$. In the following we shall need especially the validity of Cauchy's theorem and of Cauchy's formulas and the possibility of expanding a holomorphic function in power series.[4a]

If C is a closed contour satisfying the usual restrictions and if $U(\zeta)$ is holomorphic inside and on C, then $\int_C L[U(\zeta)x]\, d\zeta = 0$ for every x and L, i.e.,

$$L\left\{\left[\int_C U(\zeta)\, d\zeta\right]x\right\} = 0,$$

whence we conclude that

$$\int_C [U(\zeta)x]d\zeta = 0 \text{ and } \int_C U(\zeta)\, d\zeta = 0. \tag{2.2.5}$$

Thus Cauchy's theorem holds for holomorphic operators.

The same type of argument shows that

$$U^{(n)}(\zeta_0) = \frac{n!}{2\pi i}\int_C \frac{U(\zeta)\, d\zeta}{(\zeta - \zeta_0)^{n+1}}, \qquad n = 0, 1, 2, \cdots, \tag{2.2.6}$$

[4a] [Added in proof:] N. WIENER [1] seems to have been the first to consider analytical function theory in the strong sense in a complex space (B). For recent results and literature, consult A. E. Taylor [1].

so that Cauchy's formulas apply. Using (2.2.4) we see that if $U(\zeta)$ is holomorphic in $|\zeta - \zeta_0| \leqq \rho$ then

$$\begin{aligned} \| U^{(n)}(\zeta_0) \| &\leqq n!\rho^{-n} \frac{1}{2\pi} \int_0^{2\pi} \| U(\zeta_0 + \rho e^{i\theta}) \| \, d\theta \\ &\leqq n!\rho^{-n} \sup_{0 \leqq \theta < 2\pi} \| U(\zeta_0 + \rho e^{i\theta}) \| . \end{aligned} \tag{2.2.7}$$

For $n = 0$ this inequality shows that $\| U(\zeta) \|$ is a subharmonic function in the domain of holomorphism of $U(\zeta)$.

For $|\zeta - \zeta_0| < \rho$ we have

$$U(\zeta) = \sum_{n=0}^{\infty} \frac{1}{n!} U^{(n)}(\zeta_0)(\zeta - \zeta_0)^n \tag{2.2.8}$$

in the sense of uniform convergence since the estimates (2.2.7) show that the sum of the norms of the terms converges for such values. A similar argument applies to the series of Laurent.

2.3. Operational Binomial Series

In chapter 3 of this paper steady use will be made of expansions of operator functions in binomial series. The main properties of such series will be listed in the present paragraph.[5]

Consider the series

$$U(\zeta; x, L) \equiv \sum_{n=0}^{\infty} L(U_n x) \binom{\zeta}{n}, \qquad U_n \in \mathfrak{S}, \tag{2.3.1}$$

for fixed x and L. The series has an abscissa of convergence $\sigma_0(x, L)$ given by

$$\sigma_0(x, L) = -1 + \limsup_{n \to \infty} \log |L(V_n x)| / \log (n + 1), \tag{2.3.2}$$

where

$$V_n = U_0 - U_1 + U_2 - \cdots + (-1)^n U_n ,$$

provided the abscissa exceeds -1. This is the only case of interest to us. It is well known that the series defines an analytic function of ζ, holomorphic in the half-plane of convergence $\Re(\zeta) > \sigma_0(x, L)$. Let

$$\sigma_0 = \sup \sigma_0(x, L) \quad \text{when} \quad \| x \| = 1, \| L \| = 1.$$

We suppose that $\sigma_0 < \infty$. It follows that the series

$$\sum_{n=0}^{\infty} U_n \binom{\zeta}{n} \tag{2.3.3}$$

converges weakly for $\Re(\zeta) > \sigma_0$. If it has a weak limit $U(\zeta)$ in this domain, $U(\zeta)$ is holomorphic for $\Re(\zeta) > \sigma_0$. We leave this question open, but show the

[5] For the general theory of binomial series, also known as binomial coefficient series or Newton's interpolation series, the reader is referred to N. E. Nörlund [2]. The basic memoirs in the field are those of F. Carlson [2] and Nörlund [1].

existence of a uniform limit for $\Re(\zeta) > \sigma_0 + 1$, so that the series defines a holomorphic operator function in $\mathfrak{E}$ at least for $\Re(\zeta) > \sigma_0 + 1$. We call this function $U(\zeta)$.

For this purpose we shall estimate the coefficients. It follows from (2.3.2) that to every $\sigma > \sigma_0$ we can find a constant $M = M(\sigma, x, L)$ such that

$$| L(V_n x) | \leqq M(\sigma, x, L) \frac{\Gamma(n + \sigma + 2)}{\Gamma(\sigma + 2)\Gamma(n + 1)}$$

and the same estimate is clearly true for $| L(U_n x) |$. Using Banach's theorems quoted above, we conclude that

$$\| U_n \| \leqq B(\sigma) \frac{\Gamma(n + \sigma + 2)}{\Gamma(\sigma + 2)\Gamma(n + 1)}. \tag{2.3.4}$$

Consequently

$$\| U(\zeta) \| \leqq B(\sigma) \sum_{n=0}^{\infty} \frac{\Gamma(n + \sigma + 2)}{\Gamma(\sigma + 2)\Gamma(n + 1)} \left| \binom{\zeta}{n} \right| \tag{2.3.5}$$

and this series converges for $\Re(\zeta) > \sigma + 1$. Hence the series (2.3.3) converges strongly and uniformly at least for $\Re(\zeta) > \sigma + 1$.

The expansion of a numerical function in binomial series is not unique owing to the existence of null series. If m is an integer $\geqq 0$, the series

$$\psi_m(\zeta) = \sum_{n=m}^{\infty} (-1)^{n+m} \binom{n}{m} \binom{\zeta}{n}$$

converges for $\Re(\zeta) > m$ to the sum zero. It has the same sum for $\zeta = 0, 1, \cdots, m - 1$, and equals 1 for $\zeta = m$. This fact can be used to obtain the so called reduced series for $U(\zeta)$. We shall suppose that $U(\zeta)$ is known in advance to be holomorphic in the right half-plane and to be defined at the origin, $U(0) \in \mathfrak{E}$. We can then find coefficients $c_0, c_1, \cdots, c_p$, where $p = [\sigma_0]$ so that the function

$$U(\zeta) + \sum_{m=0}^{p} c_m \psi_m(\zeta) U(m)$$

has a binomial expansion weakly convergent for $\Re(\zeta) > \sigma_0$ and takes on the value $U(m)$ for $\zeta = m$, $m = 0, 1, \cdots, p$. The resulting binomial series is known as the reduced series. We shall always suppose that this process has been carried out in advance, so that the binomial series with which we are dealing are reduced series.

In the case of a reduced series, the coefficients are uniquely determined by the values of $U(\zeta)$ at the non-negative integers. We find by substitution

$$U_n = \sum_{k=0}^{n} (-1)^{n-k} \binom{n}{k} U(k), \qquad n = 0, 1, 2, \cdots. \tag{2.3.6}$$

Let us return to (2.3.5). Using an estimate due to F. Carlson [2] we find that for every $\beta > \sigma_0 + 1$ and for $-\pi/2 \leqq \theta \leqq \pi/2$

$$\| U(\beta + re^{i\theta}) \| \leqq B(\beta) \exp [rl(\theta)] r^{\beta + \frac{1}{2}} (1 + r \cos \theta)^{-\frac{1}{2}}, \tag{2.3.7}$$

where

$$l(\theta) = \theta \sin \theta + \cos \theta \log (2 \cos \theta). \tag{2.3.8}$$

Carlson has shown, conversely, that if an analytic function $f(\zeta)$ is holomorphic for $\Re(\zeta) \geqq \beta$ and

$$|f(\beta + re^{i\theta})| \leqq M \exp [rl(\theta)] (1 + r)^a, \tag{2.3.9}$$

then $f(\zeta)$ can be represented by a binomial series whose abscissa of convergence is less than or equal to the larger of the two numbers a and β. From this result we get immediately the following

THEOREM 2. *Let $U(\zeta)$ be an operator function holomorphic for $\Re(\zeta) > 0$ and having a value $U(0) \in \mathfrak{E}$ at the origin. Further, for every $\beta > 0$ let*

$$\| U(\beta + re^{i\theta}) \| \leqq M(\beta) \exp [rl(\theta)] (1 + r)^a. \tag{2.3.10}$$

Then

$$U(\zeta) = \sum_{n=0}^{\infty} \left\{ \sum_{k=0}^{n} (-1)^{n-k} \binom{n}{k} U(k) \right\} \binom{\zeta}{n}, \tag{2.3.11}$$

where the series is weakly convergent at least for $\Re(\zeta)$ greater than the larger of the numbers 0 and a, and is strongly as well as uniformly convergent at least for $\Re(\zeta)$ greater than the larger of the numbers 0 and $1 + a$.[6]

PROOF. We apply Carlson's theorem to the function $L[U(\zeta)x]$ with an arbitrary $x \in E$, $L \in \bar{E}$, and conclude that the series (2.3.11) is weakly convergent in the half-plane stated in the theorem. We have then merely to apply the argument which led to the estimate (2.3.5) above in order to complete the proof.

A numerical binomial series can be summed by Cesàro's method of summation, but Nörlund[7] has given a much more powerful method which applies in the largest half-plane in which the function represented by the series is of finite exponential order. If $f(\zeta)$ is holomorphic for $\Re(\zeta) \geqq \beta$, and

$$|f(\beta + re^{i\theta})| \leqq A \exp [Br], \qquad -\frac{\pi}{2} \leqq \theta \leqq \frac{\pi}{2},$$

then we can clearly determine ω so small that $f(\omega\zeta)$ satisfies (2.3.9) and is consequently representable by a convergent binomial series in the variable ζ. This gives a convergent binomial series in the variable ζ/ω for the original function $f(\zeta)$. Since $\min l(\theta) = \log 2$ it is sufficient to take $\omega < (\log 2)/B$. In particular, if $f(\zeta)$ is representable by an ordinary binomial series in the variable ζ convergent for $\Re(\zeta) > \sigma_0$, it also admits of a representation in terms of a binomial series in the variable ζ/ω for $0 < \omega < 1$. Let the abscissa of convergence of the latter

---

[6] The reader should keep in mind that the statement "uniformly convergent" refers to the topology of $\mathfrak{E}$ and not to the ζ-plane. Actually the uniform convergence holds uniformly with respect to ζ in any finite portion of the half-plane $\Re(\zeta) \geqq \max(0, 1 + a) + \epsilon$.

[7] For the following discussion compare Nörlund [1], pp. 37–44.

series be denoted by $\sigma_0(\omega)$. Nörlund has given the following method for the determination of $\sigma_0(\omega)$. Define

$$\gamma(\xi) = \limsup_{|\eta|\to\infty} \frac{1}{|\eta|} \log |f(\xi + i\eta)|.$$

There exists a largest value $\beta_0 \leqq \beta$ such that $f(\xi + i\eta)$ is holomorphic and $\gamma(\xi)$ is finite for $\xi > \beta_0$. Then $\gamma(\xi)$ is non-negative, monotone decreasing, continuous, and convex for $\xi > \beta_0$. We can find the largest interval (ω_0, ω_1) such that for $0 \leqq \omega_0 < \omega < \omega_1 \leqq +\infty$ the equation

$$\gamma(\sigma) = \frac{\pi}{2\omega}$$

has a unique solution, $\sigma = \sigma_0(\omega)$. This is the required abscissa of convergence. For $0 < \omega \leqq \omega_0$ we have $\sigma_0(\omega) = \sigma_0(\omega_0)$, for $\omega > \omega_1$, $\sigma_0(\omega) = +\infty$. Further $\sigma_0(\omega)$ is monotone increasing and continuous for $\omega_0 < \omega < \omega_1$.

Applying this method to operator functions we are led to the following theorem.

THEOREM 3. *Let $U(\zeta)$ be an operator function belonging to $\mathfrak{S}$, holomorphic for $\Re(\zeta) > 0$, and $U(0) \in \mathfrak{S}$. Suppose that for every $\delta > 0$, $\| U(\delta + re^{i\theta}) \| \leqq A_\delta \exp [B_\delta r]$, and put*

$$\gamma(\xi) = \limsup_{|\eta|\to\infty} \frac{1}{|\eta|} \log \| U(\xi + i\eta)\|, \qquad \xi > 0. \tag{2.3.12}$$

Then $\gamma(\xi)$ is non-negative, monotone decreasing, continuous, and convex for $\xi > 0$. Determine the largest interval (ω_0, ω_1) so that the equation

$$\gamma(\sigma) = \frac{\pi}{2\omega}$$

has a unique solution, $\sigma = \sigma_0(\omega)$, for $\omega_0 < \omega < \omega_1$. $\sigma_0(\omega)$ is monotone increasing and tends to zero when $\omega \to \omega_0$. Further,

$$U(\zeta) = \sum_{n=0}^{\infty} \left\{ \sum_{k=0}^{n} (-1)^{n-k} \binom{n}{k} U(k\omega) \right\} \binom{\zeta/\omega}{n}, \tag{2.3.13}$$

where the series is weakly convergent for $\Re(\zeta) > 0$ when $0 < \omega \leqq \omega_0$, and for $\Re(\zeta) > \sigma_0(\omega)$ when $\omega_0 < \omega < \omega_1$, and fails to converge weakly when $\omega > \omega_1$. The abscissas of strong and of uniform convergence exceed the abscissa of weak convergence by at most ω.

PROOF. We have merely to apply Nörlund's theorem quoted above to the family of functions $L[U(\zeta)x]$, $\| L \| = 1$, $\| x \| = 1$. Each of these functions gives rise to a growth function $\gamma(\xi; x, L)$ and an abscissa of convergence $\sigma_0(\omega; x, L)$. Further

$$\gamma(\xi) = \sup \gamma(\xi; x, L), \qquad \sigma_0(\omega) = \sup \sigma_0(\omega; x, L).$$

From these relations follow both the analytical properties of $\gamma(\xi)$ and $\sigma_0(\omega)$ and the properties of weak convergence stated in the theorem. The position of the other abscissas follows from the estimate of the coefficients which is obtained from the knowledge of the abscissa of weak convergence.

2.4. Semi-Groups of Transformations

After these preliminaries we are prepared to take up the study of semi-groups.

Let $\mathfrak{S} = \{T_\alpha\}$ be a one-parameter family of transformations having the following properties:

(T_1) *$T_\alpha \in \mathfrak{S}$ for $\alpha \geqq 0$ and T_0 is the identity.*

(T_2) *For each x in E we have*

$$T_\alpha[T_\beta x] = T_\beta[T_\alpha x] = T_{\alpha+\beta}x, \qquad \alpha \geqq 0, \qquad \beta \geqq 0. \tag{2.4.1}$$

Under these circumstances we call $\mathfrak{S}$ a one-parameter *semi-group*. It is evident that a semi-group has several properties in common with Abelian groups. Thus there exists a unit-element, the identical transformation. Further, the composition, in our case addition, is unrestricted, associative, and commutative. But ordinarily the elements of $\mathfrak{S}$ do not have inverses, that is unless T_α produces a one-to-one map of E upon itself.

We have supposed here that T_α is defined for $0 \leqq \alpha < \infty$. If we suppose instead that T_α is defined merely for $0 \leqq \alpha \leqq a$ and that (2.4.1) holds whenever α, β, and $\alpha + \beta$ belong to the interval $[0, a]$, then we can extend $\{T_\alpha\}$ to become a semi-group. Indeed, it is evidently possible to define transformations T_α for $\alpha > a$ with the aid of (2.4.1) in such a manner that the resulting extended set forms a semi-group.

A different situation presents itself when T_α is defined for $0 < a \leqq \alpha \leqq b < \infty$ and (2.4.1) holds when α, β, and $\alpha + \beta$ belong to $[a, b]$. The latter assumption presupposes that $b > 2a$. Here an extension for $\alpha < a$ or $b < \alpha$ does not appear to be always possible (cf. §2.5). We shall call such a set $\{T_\alpha\}$ an *incomplete semi-group*.

Let $\mathfrak{S}$ be a semi-group. We say that $\mathfrak{S}$ is *weakly measurable* when the operator function T_α is weakly measurable for $\alpha \geqq 0$. The property of being weakly measurable is not implied by the semi-group property. Indeed, let $f(u)$ be a real, non-measurable solution[8] of the equation $f(u + v) = f(u) + f(v)$. Then

$$T_\alpha x = e^{if(\alpha)}x, \qquad \alpha \geqq 0, \tag{2.4.2}$$

defines a semi-group which obviously is not weakly measurable. Since $\| T_\alpha x \| = \| x \|$, $\| T_\alpha \| = 1$, we conclude that measurability, or even continuity and analyticity, of $\| T_\alpha \|$ does not imply the weak measurability of a semi-group.

We say that $\mathfrak{S}$ is *bounded on a finite interval* (a, b), $0 \leqq a < b < \infty$, if

$$\| T_\alpha \| \leqq M, \qquad a \leqq \alpha \leqq b. \tag{2.4.3}$$

[8] Such solutions exist according to Hamel [1].

It is *bounded on an infinite interval* (a, ∞) if

$$\| T_\alpha \| \leqq Me^{B\alpha}, \qquad a \leqq \alpha < \infty. \tag{2.4.4}$$

It is clear that (2.4.4) implies (2.4.3), but the converse is also true by virtue of the semi-group property, the resulting value of B being $\leqq (\log M)/b$. In particular, the inequality

$$\limsup_{\alpha \to 0} \| T_\alpha \| = A < \infty \tag{2.4.5}$$

implies the boundedness of $\mathfrak{S}$ on $(0, \epsilon)$ and hence also on $(0, \infty)$.

The property of being bounded is not implied by the semi-group property. Indeed, suppress the factor i in the exponent in formula (2.4.2) and the result is a non-measurable semi-group which is not bounded in any interval. Whereas boundedness does not imply measurability, measurability of the norm implies at least some types of boundedness.

THEOREM 4. *If* $\| T_\alpha \|$ *is measurable for* $\alpha > 0$, *then* $\| T_\alpha \|$ *is bounded on any interval* $[a, b]$, $0 < a < b < \infty$.

PROOF. Put

$$e(\alpha) = \log \| T_\alpha \|, \qquad p(\alpha) = \tfrac{1}{2}[e(\alpha) + | e(\alpha) |]. \tag{2.4.6}$$

By assumption these are measurable functions of α. From the semi-group property follows that

$$e(\alpha + \beta) \leqq e(\alpha) + e(\beta), \tag{2.4.7}$$

and a simple argument shows that the same inequality is satisfied by $p(\alpha)$. We want to show that $p(\alpha)$ is bounded on $[a, b]$, $a > 0$. Suppose contrariwise. Then there exists a $\gamma \geqq a$ and a sequence $\alpha_n \to \gamma$ such that $p(\alpha_n) \geqq 2n$. Now let α and β be arbitrary positive numbers whose sum equals α_n, then $p(\alpha) + p(\beta) \geqq 2n$ by (2.4.7). This implies that the subset on $[0, \alpha_n]$ where $p(\alpha) \geqq n$ has a measure $\geqq \alpha_n/2$ for every n. Consequently $p(\alpha) = + \infty$ on a set of measure $\geqq a/2$. But this is impossible, because $T_\alpha \epsilon \mathfrak{S}$, i.e., $\| T_\alpha \|$ has a finite definite value for every fixed $\alpha > 0$. It follows that $p(\alpha)$ is bounded in $[a, b]$ and consequently also $\| T_\alpha \|$.

The argument does not permit us to show that measurability of $\| T_\alpha \|$ implies (2.4.5), i.e., boundedness of $\mathfrak{S}$ on $(0, \infty)$. For this question see §5.2.

We have already defined various types of continuity for operator functions. Suppose that $\mathfrak{S}$ is a semi-group and that T_α is continuous on the right in the weak sense at $\alpha = \alpha_0 \geqq 0$, i.e., for every $x \epsilon E$, $L \epsilon \bar{E}$,

$$L(T_{\alpha_0 + h} x) \to L(T_{\alpha_0} x) \quad \text{as} \quad h \downarrow 0.$$

Let $\alpha > \alpha_0$. Then

$$L(T_{\alpha+h} x) = L[T_{\alpha_0+h}(T_{\alpha-\alpha_0} x)] \to L[T_{\alpha_0}(T_{\alpha-\alpha_0} x)] = L(T_\alpha x),$$

i.e., continuity on the right at $\alpha = \alpha_0$ implies the same property for $\alpha > \alpha_0$.

The semi-group property alone does not imply any kind of continuity of T_α

as is obvious from our previous examples involving non-measurable semi-groups. But N. Dunford [1] has proved the following

THEOREM. *If $\mathfrak{S}$ is bounded on $(0, \infty)$ and weakly measurable and if, in addition, for every $x \in E$ the set $\{T_\alpha x\}$, $\alpha > 0$, is separable, in particular, if E is separable, then T_α is strongly continuous on the right for $\alpha > 0$. If the set $\{T_\alpha\}$ of points in $\mathfrak{S}$ is separable in the uniform topology of $\mathfrak{S}$, then T_α is uniformly continuous on the right for $\alpha > 0$.*

Finally we note that continuity of a semi-group does not imply differentiability. This is shown by the following example. Let E be the Lebesgue space $L_2(-\infty, \infty)$ with the customary metric and put

$$T_\alpha x = e^{i\alpha t}x(t), \qquad \alpha \geqq 0.$$

Then T_α generates a semi-group $\mathfrak{S}$ which is uniformly continuous. But if we form the difference quotients

$$\frac{1}{h}[T_{\alpha+h} - T_\alpha]x = \frac{1}{h}\{e^{ih t} - 1\}e^{i\alpha t}x(t),$$

we have a sequence of transformations the norms of which tend to infinity as $h \to 0$. It follows that T_α is not differentiable even in the weak sense in spite of the fact that $T_\alpha x$ is an analytic function of α for fixed x and t.

2.5. Analytic Extension and the Semi-Group Property

So far we have only considered semi-groups defined on the positive real axis. Suppose now that we have an operator function $W(\zeta)$ defined in a domain Z_0 of the complex plane and such that

$$W(\zeta_1)[W(\zeta_2)] = W(\zeta_2)[W(\zeta_1)] = W(\zeta_1 + \zeta_2), \tag{2.5.1}$$

whenever ζ_1, ζ_2, and $\zeta_1 + \zeta_2$ belong to Z_0. We then say that $W(\zeta)$ has the *semi-group property in Z_0*. *If*, in particular, *Z_0 is closed under addition*, i.e., $\zeta_1 + \zeta_2$ belongs to Z_0 whenever ζ_1 and ζ_2 do so, *and if*, in addition, *$\zeta = 0$ is a boundary point of Z_0 and $W(0) = I$, we say that $W(\zeta)$ defines a semi-group in Z_0*.

If a semi-group $\mathfrak{S} = \{T_\alpha\}$ be given for $\alpha \geqq 0$, it is always possible to embed it in a semi-group defined for the right half-plane. It is sufficient to take $W(\xi + i\eta) = e^{i\eta}T_\xi$. This extension is obviously not analytic except in the special case in which $T_\xi = e^\xi$.

On the other hand, if an incomplete semi-group admits of an analytic extension, then the extension has the semi-group property in the domain of analyticity. Before we prove this result, let us define the *linear extension* Z of Z_0. A point ζ shall belong to Z if and only if there are points $\zeta_1, \zeta_2, \cdots, \zeta_k$ in Z_0 and positive integers $m_1, m_2, \cdots, m_k$ such that

$$\zeta = m_1\zeta_1 + m_2\zeta_2 + \cdots + m_k\zeta_k.$$

We shall prove the following

THEOREM 5. *Let* $\{T_\alpha\}$, $0 < a < \alpha < b < \infty$ $(2a < b)$ *be an incomplete semi-group. Let* $W(\zeta)$ *be an operator function defined in a domain* Z_0 *of the complex* ζ*-plane, containing the segment* (a, b) *of the real axis, let* $W(\zeta)$ *be holomorphic in* Z_0 *and* $W(\xi) = T_\xi$, $a < \xi < b$. *Then* $W(\zeta)$ *has the semi-group property in* Z_0. *Moreover,* $W(\zeta)$ *can be defined as a holomorphic function having the semi-group property in the linear extension* Z *of* Z_0.

PROOF. Let α be a fixed real number, $a < \alpha < b/2$, and form the two operator functions

$$W(\zeta)[W(\alpha)] \quad \text{and} \quad W(\alpha)[W(\zeta)].$$

They exist and are holomorphic functions of ζ in Z_0. For $\zeta = \xi$ real, $a < \xi < b/2$, they coincide, the common value being $W(\xi + \alpha)$. They are consequently identical within their common domain of analyticity and coincide with $W(\zeta + \alpha)$ as long as also $\zeta + \alpha \in Z_0$. Let $Z_0(\alpha)$ be the domain $Z_0 + \alpha$, i.e., $\zeta \in Z_0(\alpha)$ if $\zeta - \alpha \in Z_0$.

In that portion of $Z_0(\alpha)$ which is outside of Z_0 we define

$$W(\zeta) = W(\zeta - \alpha)[W(\alpha)]. \tag{2.5.2}$$

This definition makes sense since $\zeta - \alpha \in Z_0$. It defines $W(\zeta)$ as a holomorphic function of ζ in $Z_0(\alpha)$. Suppose now that $\zeta \in Z_0(\alpha) \cdot Z_0(\beta)$, where $a < \alpha < \beta < b/2$. Since

$$W(\xi - \alpha)[W(\alpha)] = W(\xi - \beta)[W(\beta)]$$

for $a + \beta < \xi < b/2 + \alpha$, we conclude that

$$W(\zeta - \alpha)[W(\alpha)] = W(\zeta - \beta)[W(\beta)], \qquad \zeta \in Z_0(\alpha) \cdot Z_0(\beta).$$

Hence the definition of $W(\zeta)$ is unique in the union of all sets $Z_0(\alpha)$, $a < \alpha < b/2$, and is independent of α.

Now consider the operator function

$$W(\zeta - \alpha)[W(\alpha)].$$

This function is defined for $\alpha \in Z_0$, $\zeta - \alpha \in Z_0$, and is a holomorphic function of α as well as of ζ in the domain of definition. Let $a < \alpha_0 < b/2$ and consider a small circular region $\gamma_1 : |\alpha - \alpha_0| < \rho$. Let ζ_0 be such that $\zeta_0 - \alpha_0 \in Z_0$ and consider the circular region $\gamma_2 : |\zeta - \zeta_0| < \rho$. If ρ is suitably chosen, all points of γ_1 are in Z_0 and all points of γ_2 are in $Z_0(\alpha_0)$. When α is real and in γ_1, and $\zeta \in \gamma_2$, then $W(\zeta - \alpha)[W(\alpha)] = W(\zeta)$, i.e., is independent of α. It follows that this relation holds everywhere in γ_1 and hence throughout the domain of analyticity of $W(\zeta)$. Thus the definition

$$W(\zeta) = W(\zeta - \alpha)[W(\alpha)]$$

holds for $\alpha \in Z_0$, $\zeta - \alpha \in Z_0$. It defines $W(\zeta)$ uniquely as a holomorphic operator function satisfying the semi-group property. In particular, the semi-group property is found to hold in Z_0.

At this stage of the process the function $W(\zeta)$ is defined for ζ in Z_1, where Z_1 is the set of all points ζ of the form $\zeta = m_1\zeta_1 + m_2\zeta_2$ with ζ_1 and ζ_2 in Z_0, m_1 and m_2 being non-negative integers, $1 \leqq m_1 + m_2 \leqq 2$. In particular, the line-segment $a < \xi < 2b$ belongs to Z_1. We can now apply exactly the same argument and the same process of extension to the domain Z_1 as we have just applied to Z_0. This leads to the definition of $W(\zeta)$ as a holomorphic function satisfying the semi-group property in a domain Z_2. Here Z_2 contains all points of the form $\zeta = \sum_1^4 m_k\zeta_k$ where ζ_k's are in Z_0 and the m_k's are non-negative integers having a sum between one and four. The procedure can then be repeated indefinitely. It is clear that every point of Z, the linear extension of Z_0 defined above, is reached after a finite number of extensions and belongs to the domain of holomorphism of $W(\zeta)$. Further, the semi-group property holds in Z by virtue of the process of extension. This completes the proof of the theorem. We have assumed tacitly that Z_0 is not its own linear extension in which case of course the last part of the proof becomes superfluous.

We see in particular that the line-segment (a, ∞) belongs to Z, i.e., an incomplete semi-group defined on $a < \xi < b$ $(2a < b)$ can always be extended for $\xi \geqq b$ in such a manner that the semi-group property is preserved, provided T_ξ is analytic on the line-segment $a < \xi < b$. It would seem likely that in this statement analyticity can be replaced by quasi-analyticity.

Theorem 5 shows that the domain of holomorphism Z of an operator function $W(\zeta)$, which defines a semi-group, is closed under addition. If it contains a point ζ_0 it must contain all points of the form $\zeta_0 + \zeta_1$ where $\zeta_1 \epsilon Z$. In the simplest case, that in which Z contains the positive real axis, we know that all points $\zeta_0 + h, h > 0$, are in Z. In any case we can find a sequence $\{\zeta_n\}$ such that $\zeta_n \epsilon Z$, $|\zeta_n| > |\zeta_{n+1}|$, $\zeta_n \to 0$, and $\arg \zeta_n$ tends to a finite limit. If the distance of ζ_0 from the boundary of Z is supposed to be δ, then all points ζ belong to Z which have a distance less than δ from any one of the points $\zeta_0 + k\zeta_n$, where $k = 0, 1, 2, \cdots$, $n = 1, 2, 3, \cdots$. In particular the rays $\arg(\zeta - \zeta_0) = \arg \zeta_n$ belong to Z for n sufficiently large. We conclude from these facts that Z *is simply connected.*

It is obvious that Z contains sectors of the form $\theta_1 < \arg(\zeta - \zeta_0) < \theta_2$. If $\theta_2 - \theta_1$ can be taken greater than π, it is readily seen that Z must be the whole plane. In this case $W(\zeta)$ defines a group rather than a semi-group. Let us exclude this case. Z is then contained in some half-plane $\theta_1 < \arg \zeta < \theta_1 + \pi$, and it is no restriction to assume $\theta_1 = -\pi/2$, so that Z is some subset of the right half-plane. We see thus that the right half-plane can be regarded as the maximum domain of holomorphism of a semi-group.

The boundary Γ of Z is the natural boundary of $W(\zeta)$ in the following sense. To every point ζ_0 on Γ we can find elements $x \epsilon E$, $L \epsilon \bar{E}$ such that $L[W(\zeta)x]$ admits of ζ_0 as a singular point. Judging from simple examples it should be possible to determine x and L so that $L[W(\zeta)x]$ has Γ as its natural boundary and it may even be possible to determine an L to every x so that this is the case. I am not able to prove these surmises, however.

We shall say that a family of operators $\{W(\zeta)\}$ having the semi-group property in a domain Δ is bounded in Δ if there exist constants A and B such that

$$\| W(\zeta) \| \leqq A \exp [B | \zeta |], \qquad \zeta \in \Delta. \tag{2.5.3}$$

Let us consider the most important case, that in which $\{W(\zeta)\}$ forms a semi-group in the right half-plane, i.e., the maximal domain of existence. Let S_ϵ be the sector $\epsilon \leqq | \zeta |$, $| \arg \zeta | \leqq \pi/2 - \epsilon$, $\epsilon > 0$. It is readily seen that $\mathfrak{S}$ is bounded in S_ϵ for every $\epsilon > 0$ when $W(\zeta)$ is holomorphic. Indeed, if this were not the case, then there would exist a monotone increasing function $\omega(r) \to \infty$ with r and a sequence of points $\zeta_n \in S_\epsilon$, where $| \zeta_n | = r_n \to \infty$ such that

$$\| W(\zeta_n) \| \geqq \exp [r_n \omega(r_n)].$$

But

$$\| W(\zeta/m) \|^m \geqq \| W(\zeta) \| .$$

Hence choosing $\zeta = \zeta_n$, $m = [r_n]$, we get

$$\| W(\zeta_n/[r_n]) \| \geqq \exp (\omega(r_n))$$

and this would imply that $\| W(\zeta) \|$ would not be bounded in the domain $1 < | \zeta | < 1 + \delta$, $| \arg \zeta | < \pi/2 - \epsilon$. This contradicts the assumption that $W(\zeta)$ is holomorphic for $\xi > 0$. Here we conclude that $\mathfrak{S}$ is bounded in S_ϵ for every $\epsilon > 0$.

On the other hand, it does not seem to follow from the semi-group property that $W(\zeta)$ is necessarily bounded in any right half-plane. Nor is the rate of growth of $\| W(\xi + i\eta) \|$ for fixed ξ as $| \eta | \to \infty$ or for fixed η as $\xi \downarrow 0$ subject to any obvious restrictions.

3. Analytical Semi-Groups Bounded in a Half-Plane

3.1. Necessary Conditions

Let the operator function $W(\zeta)$ define a semi-group bounded and holomorphic in every half-plane $\xi \geqq \delta > 0$. We shall characterize $W(\zeta)$ by means of the properties of its resolvent for small real values of ζ.

Theorem 3 shows that $W(\zeta)$ can be represented by a binomial series in the variable ζ/ω for sufficiently small values of ω. Since

$$\sum_{k=0}^{\infty} (-1)^k \binom{n}{k} W(k\omega) = [I - W(\omega)]^n, \tag{3.1.1}$$

the series takes the form

$$W(\zeta) = \sum_{k=0}^{\infty} (-1)^n [I - W(\omega)]^n \binom{\zeta/\omega}{n} \tag{3.1.2}$$

or symbolically

$$W(\zeta) = \{I - [I - W(\omega)]\}^{\zeta/\omega} = [W(\omega)]^{\zeta/\omega}. \tag{3.3.3}$$

Here, however, we have to define the power with the aid of the preceding series when the latter is convergent.

Let the abscissa of weak convergence be $\sigma_0(\omega)$. It is determined by the method given in Theorem 3. We know that $\sigma_0(\omega) \downarrow 0$ as $\omega \downarrow 0$. If $\omega < \omega_1$ and $\sigma > \sigma_0(\omega)$, formula (2.3.4) shows that

$$(3.1.4) \qquad \| [I - W(\omega)]^n \| \leqq B(\sigma, \omega) \frac{\Gamma(\sigma/\omega + n + 2)}{\Gamma(\sigma/\omega + 2)\Gamma(n + 1)}.$$

Let us now consider the series

$$(3.1.5) \qquad -\sum_{n=1}^{\infty} (1 - \lambda)^{-n-1}[I - W(\omega)]^n \equiv R(\lambda, \omega).$$

We have

$$(3.1.6) \qquad \| R(\lambda, \omega)\| \leqq B(\sigma, \omega) \left\{1 - \frac{1}{|\lambda - 1|}\right\}^{-2-\sigma/\omega}$$

by (3.1.3). It follows that the series converges in the sense of the uniform topology for $|\lambda - 1| > 1$.

Following Hadamard we define the *order of an operator* power series $\sum_0^\infty U_n\zeta^n$ on its circle of convergence, which we take to have the radius one, as

$$1 + \limsup_{n\to\infty} \log \| U_n \|/\log n.$$

Using (3.1.4) we see that $R(\lambda, \omega)$ is of finite order $\rho(\omega)$ on the circle $|\lambda - 1| = 1$ and

$$(3.1.7) \qquad \rho(\omega) \leqq \frac{1}{\omega} \sigma_0(\omega) + 2.$$

A simple calculation shows that for $|\lambda - 1| > 1$ we have

$$(3.1.8) \qquad R(\lambda, \omega)[\lambda I - W(\omega)] = [\lambda I - W(\omega)]R(\lambda, \omega) = I.$$

In other words, $R(\lambda, \omega)$ is the *resolvent* of the transformation $W(\omega)$ and this resolvent exists as a bounded transformation in $\mathfrak{E}$ outside of the circle $|\lambda - 1| = 1$ for $0 < \omega < \omega_1$ and is of finite order on the circle.[9] Thus we have proved the following theorem.

THEOREM 6. *Let $W(\zeta)$ be holomorphic and define a semi-group in the right half-plane and let $W(\zeta)$ be bounded in the sense of formula* (2.5.3) *in every half-plane $\xi \geqq \delta > 0$. Let*

$$\gamma(\xi) = \limsup_{|\eta|\to\infty} \frac{1}{|\eta|} \log \| W(\xi + i\eta)\|, \qquad \xi > 0,$$

[9] For the definition and properties of the resolvent see M. H. Stone [1], pp. 128–141. The extension from Hilbert space to a general space of type (B) does not cause any difficulties.

and put $\gamma(\infty) = \lim_{\xi\to\infty}\gamma(\xi)$. *Finally let*

$$\omega_1 = \frac{\pi}{2\gamma(\infty)}.$$

Then for $0 < \omega < \omega_1$ *the resolvent* $R(\lambda, \omega)$ *exists and is holomorphic for* $|\lambda - 1| > 1$. *Moreover,* $R(\lambda, \omega)$ *is of finite order* $\rho(\omega) \leqq (1/\omega)\sigma_0(\omega) + 2$ *on the circle* $|\lambda - 1| = 1$ *and satisfies* (3.1.6) *for any* $\sigma > \sigma_0(\omega)$.

3.2. Spectral Properties of Semi-Groups, I

Suppose that $\mathfrak{S} : \{T_\alpha\}$ defines a semi-group on the positive real axis. In the complex λ-plane let $S(\alpha)$ and $R(\alpha)$ denote the *spectrum* and the *resolvent set* respectively of the transformation T_α. We recall that if $\lambda_1 \epsilon R(\alpha)$, then $R(\lambda_1, \alpha) \epsilon \mathfrak{S}$, whereas if $\lambda_2 \epsilon S(\alpha)$, then $R(\lambda_2, \alpha)$ may cease to exist altogether and in any case does not belong to $\mathfrak{S}$.

We have obviously

$$R(\lambda^n, n\alpha) = \prod_{k=1}^{n} R(\epsilon^k\lambda, \alpha), \qquad \epsilon = \exp(2\pi i/n), \tag{3.2.1}$$

$$R(\lambda, \alpha) = R(\lambda^n, n\alpha)\sum_{j=0}^{n-1}\lambda^j T_{j\alpha}, \tag{3.2.2}$$

when the resolvents involved belong to $\mathfrak{S}$. The first formula shows that if all the resolvents on the right hand side exist so does the left hand side; hence if the left side does not exist in $\mathfrak{S}$, at least one factor on the right is not in $\mathfrak{S}$. The second formula shows that the existence of $R(\lambda^n, n\alpha)$ implies that of $R(\lambda, \alpha)$, but not vice versa. In other words, if $\lambda_1 \epsilon S(\alpha)$ then $\lambda_1^n \epsilon S(n\alpha)$ and at least one of the n^{th} roots of λ_1 belongs to $S(\alpha/n)$, whereas if $\lambda_2 \epsilon R(\alpha)$ then all the n^{th} roots of λ_2 belong to $R(\alpha/n)$ but λ_2^n need not belong to $R(n\alpha)$. We can express this symbolically by

$$[R(\alpha/n)]^n \supset R(\alpha), \tag{3.2.3}$$

$$[S(\alpha/n)]^n \subset S(\alpha), \tag{3.2.4}$$

where we use the symbol D^n to denote the domain of the n^{th} powers of the quantities in a given domain D.

Some applications of these relations will be useful later. Since T_α is a bounded transformation, the series

$$\sum_{n=0}^{\infty}\lambda^{-n-1}T_{n\alpha} = R(\lambda, \alpha) \tag{3.2.5}$$

converges uniformly for $|\lambda| > \|T_\alpha\|$. That its sum is $R(\lambda, \alpha)$ is verified by multiplying on the left or on the right by $\lambda I - T_\alpha$ which gives the result I in either case. Thus $S(\alpha)$ is located in the circle $|\lambda| \leqq \|T_\alpha\|$. From (3.2.4) we conclude that $S(\alpha/n)$ is located in the circle $|\lambda| \leqq \|T_\alpha\|^{1/n}$. This is better

than the obvious estimate $|\lambda| \leqq \| T_{a/n} \|$. Since $\| T_\alpha \|$ is bounded on an interval of the form $0 < a/2 \leqq \alpha \leqq a$ when $\| T_\alpha \|$ is measurable for $\alpha > 0$ by Theorem 4, we see that this hypothesis implies that $S(\alpha)$ is uniformly bounded for $0 < \alpha \leqq a$.

Theorem 6 shows that the resolvent sets corresponding to holomorphic semi-groups are subject to special restrictions. Let us suppose that $S(\alpha)$ is located in the right half-plane for $0 < \alpha \leqq a$. Formula (3.2.4) shows that this result is self-improving. In particular $S(\alpha 2^{-n})$ is located in the sector $|\arg \lambda| \leqq \pi 2^{-n-1}$, $n = 0, 1, 2, \cdots$. Indeed, for $n = 0$ this is merely our assumption. For $n = 1$ we know that $S(\alpha/2)$ can consist only of square-roots of quantities in $S(\alpha)$ and here only the principal determination of the square-root can be used. Hence $S(\alpha/2)$ is located in the sector $|\arg \lambda| \leqq \pi/4$. The proof is then accomplished by complete induction. We can also formulate the result as follows. The assumption that $S(\alpha)$ belongs to the right half-plane for $0 < \alpha \leqq a$ implies that $S(\alpha)$ belongs to the sector $|\arg \lambda| \leqq \pi 2^{-n-1}$ for $2^{-n-1}a < \alpha \leqq 2^{-n}a$.

Let us now combine this result with the previous one. Suppose that $\| T_\alpha \|$ is measurable for $\alpha > 0$ and $S(\alpha)$ is located in the right half-plane for $0 < \alpha \leqq a$. Then we can find an a_0 such that $S(\alpha)$ is located in the circle $|\lambda - 1| \leqq 1$ for $0 < \alpha \leqq a_0$. This result is also self-improving. Using formula (3.2.4) once more we obtain the following theorem.

THEOREM 7. *If $\| T_\alpha \|$ is measurable for $\alpha > 0$ and the spectrum $S(\alpha)$ of T_α is located in the right half-plane for $0 < \alpha \leqq a$ then there exists an a_0 with the following property. For $2^{-n-1}a_0 < \alpha \leqq 2^{-n}a_0$, $S(\alpha)$ is located in the domain, $\lambda = re^{i\theta}$,*

$$r^{2^n} \leqq 2 \cos(2^n\theta), \quad |\theta| \leqq \pi 2^{-n-1}, \qquad n = 0, 1, 2, \cdots. \tag{3.2.6}$$

We are also interested in the order of the resolvent on the circle $|\lambda - 1| = 1$. Theorem 7 shows that the only point that really matters is the origin. Suppose now that for a particular value of α the resolvent is of finite order at the origin. Can anything be said concerning the order of the resolvent at $\lambda = 0$ for other values of α? Theorem 8 below gives an answer to this question.

Consider a sector $V_\psi : 0 < |\lambda| \leqq 1, \psi \leqq \arg \lambda \leqq 2\pi - \psi$, where $0 < \psi < \pi/2$ so the opening of the sector is greater than π. Consider an operator function $U(\lambda)$ holomorphic in V_ψ. Put

$$M(r) = \sup \| U(re^{i\theta}) \|, \qquad \psi \leqq \theta \leqq 2\pi - \psi. \tag{3.2.7}$$

We say that $U(\lambda)$ is of *sub-exponential order* in V_ψ if

$$\liminf_{r \to 0} r^\gamma \log M(r) = 0, \qquad \pi\gamma = 2(\pi - \psi), \tag{3.2.8}$$

whereas $U(\lambda)$ is of *finite order* τ in V_ψ if

$$\limsup_{r \to 0} \log M(r) / \log \frac{1}{r} = \tau. \tag{3.2.9}$$

With this terminology agreed upon, we have the following result.

THEOREM 8. *Let $R(\lambda, \alpha)$ be holomorphic and of subexponential order in a fixed sector V_ψ for $0 < \alpha \leqq a$, and of finite order $\tau(a)$ for $\alpha = a$. Then $R(\lambda, \alpha)$ is of finite order $\tau(\alpha_n)$ for $\alpha = \alpha_n = a2^{-n}$, $n = 0, 1, 2, \cdots$, and $\alpha_n \tau(\alpha_n)$ does not increase with n.*

PROOF. Consider the relation

$$R(\sqrt{\lambda}, \alpha/2) = (\sqrt{\lambda} I + T_{\alpha/2}) R(\lambda, \alpha).$$

We know that if $\lambda \in R(\alpha)$ then $\pm\sqrt{\lambda} \in R(\alpha/2)$. Now choose $\lambda = re^{2i\psi}$, $\sqrt{\lambda} = \sqrt{r}e^{i\psi}$. A simple calculation shows that

$$\limsup_{r\to 0} \log \| R(\sqrt{r}e^{i\psi}, \alpha/2) \| / \log \frac{1}{r} \leqq \limsup_{r\to 0} \log \| R(re^{2i\psi}, \alpha) \| / \log \frac{1}{r} \leqq \tau,$$

if $R(\lambda, \alpha)$ is of finite order $\tau = \tau(\alpha)$ in V_ψ. Thus to every $\epsilon > 0$ and for $\theta = \psi$ we have

$$\| R(re^{i\theta}, \alpha/2) \| \leqq A(\alpha, \epsilon) r^{-2\tau-\epsilon},$$

and the same inequality is evidently valid for $\theta = 2\pi - \psi$.

Now form the operator function

$$\lambda^{2\tau+\epsilon} R(\lambda, \alpha/2).$$

It is holomorphic in the sector V_ψ and we have just seen that its norm is bounded on the border of the sector. Since it is of subexponential order in V_ψ, the usual Phragmén-Lindelöf argument shows that its norm is bounded in V_ψ for every $\epsilon > 0$. It follows that $R(\lambda, \alpha/2)$ is also of finite order in V_ψ and, if its order be denoted by $\tau(\alpha/2)$, we have $\tau(\alpha/2) \leqq 2\tau(\alpha)$. The proof of the theorem then follows by complete induction.

3.3. Sufficient Conditions

We are interested in the problem of finding sufficient conditions in order that a semi-group, given on the positive real axis, will admit of an analytic extension in some right half-plane. Theorem 6 gives necessary conditions for the existence of such an extension bounded in every half-plane $\xi \geqq \delta > 0$. In addition, the operator function defining the semi-group must satisfy obvious conditions of measurability, continuity, and differentiability on the real axis. We shall prove, however, that a much weaker set of conditions ensures the existence of the desired analytic extension. We shall be concerned with operator functions T_α satisfying the following six postulates:

(T_1) *and* (T_2) *as in* §2.4.

(T_3) $\| T_\alpha \|$ *is measurable for* $\alpha > 0$.

(T_4) T_α *is weakly continuous on the right at* $\alpha = \alpha_0$, *where* α_0 *is fixed,* $\geqq 0$, *but arbitrary.*

(T_5) $S(\alpha)$, *the spectrum of* T_α, *is located in the right half-plane for* $0 < \alpha \leqq a_1$.

(T_6) *$R(\lambda, \alpha)$, the resolvent of T_α, is of sub-exponential order in a fixed sector V_ψ for $0 < \alpha \leqq a_2$ and is of finite order for $\alpha = a_2$.*

These conditions will turn out to be sufficient. While they compare favorably with the necessary conditions, they are not likely to be independent or reduced to minimum terms. Various modifications are possible. Thus if the space E is separable we can replace (T_3) and (T_4) by

(T_3^*) *$\| T_\alpha \|$ is bounded on $0 < \alpha \leqq a$.*

(T_4^*) *T_α is weakly measurable for $\alpha > 0$.*

By virtue of Dunford's theorem these conditions imply (T_3) and (T_4) when E is separable. We shall now prove

THEOREM 9. *If $\mathfrak{S}:\{T_\alpha\}$ satisfies postulates (T_1) to (T_6), then there exists an operator function $W(\zeta)$ and a constant $c \geqq 0$ such that $W(\zeta)$ is holomorphic for $\xi > c$ and bounded for $\xi \geqq c + \delta$, $\delta > 0$, and $W(\xi) = T_\xi$ for $\xi \geqq c$. $W(\zeta)$ has the semi-group property in its domain of existence.*

PROOF. It is enough to prove the existence of $W(\zeta)$, the semi-group property follows from Theorem 5. We note first that (T_3) to (T_6) imply the validity of Theorems 7 and 8. Put $\alpha_n = a_2 2^{-n}$. Here we can choose n so large that $S(\alpha_n)$ is located entirely within the circle $|\lambda - 1| = 1$, except for the origin, and of course entirely outside of the sector V_ψ. For such an n we write $\alpha_n = h$ and form the integral operator

$$W(\zeta, h) = \frac{1}{2\pi i}\int_C R(\lambda, h)\lambda^{\zeta/h}\, d\lambda, \tag{3.3.1}$$

where the integral starts and ends at the origin and surrounds $S(h)$ once in the positive sense. We can integrate, for instance, along the circle $|\lambda - 1| = 1$. The power is given its principal determination. By Theorem 8 the resolvent $R(\lambda, h)$ is of finite order $\tau(h)$ at the origin. Hence the integral will exist for

$$\xi > h[\tau(h) - 1] \tag{3.3.2}$$

and defines an analytic function, holomorphic in this half-plane.[10] Since $R(\lambda, h)$ is holomorphic outside of $|\lambda - 1| = 1$ we have

$$R(\lambda, h) = -\sum_{n=0}^{\infty} (1 - \lambda)^{-n-1}(I - T_h)^n, \quad |\lambda - 1| > 1. \tag{3.3.3}$$

(Compare formula (3.1.5).) Cauchy's formula gives

$$(I - T_h)^n = \frac{1}{2\pi i}\int_C R(\lambda, h)(1 - \lambda)^n\, d\lambda,$$

where we integrate along any circle $|\lambda - 1| = \rho > 1$. Choosing $\rho = 1 + \frac{1}{n}$ and using the estimate

---

[10] The integral (3.3.1) is analogous to that of S. Pincherle [1] the existence of which is necessary and sufficient for the representability of an analytic function by means of a binomial series. The procedure in the present paragraph was suggested by pp. 43–46 of F. Carlson [2].

$$\| R(\lambda, h) \| \leqq A(h, \epsilon) \mid \lambda \mid^{-\tau(h)-\epsilon},$$

which is valid along the contour of integration, we get

$$\| (I - T_h)^n \| \leqq B(h, \epsilon) n^{\tau(h)-1+\epsilon}. \tag{3.3.4}$$

Substituting (3.3.3) into (3.3.1) and integrating termwise, which is certainly justified for ζ's with sufficiently large real parts, we get

$$W(\zeta, h) = \sum_{n=0}^{\infty} (-1)^n (I - T_h)^n \binom{\zeta/h}{n}. \tag{3.3.5}$$

Formula (3.3.4) shows that this series converges in the uniform sense at least for

$$\xi > h\,[\tau(h) - 1] \equiv \sigma_h\,. \tag{3.3.6}$$

Carlson's estimate (2.3.7) shows that for $\beta > \sigma_h$ and $-\pi/2 \leqq \theta \leqq \pi/2$

$$\| W(\beta + re^{i\theta}, h) \| \leqq M(\beta, h) \exp\left[\frac{r}{h}\, l(\theta)\right] r^{\sigma_h+\frac{1}{2}+\epsilon}. \tag{3.3.7}$$

Finally we note that

$$W(kh, h) = T_{kh}\,, \tag{3.3.8}$$

where k is any integer greater than $\tau(h) - 1$.

The preceding analysis applies to any h of the form $h = \alpha_n = a_2 2^{-n}$ with sufficiently large n. Let us put

$$W_n(\zeta) = W(\zeta, \alpha_n) \tag{3.3.9}$$

and write σ_n and τ_n for $\sigma(h)$ and $\tau(h)$ when $h = \alpha_n$.

We shall prove that $W_n(\zeta)$ and $W_{n+1}(\zeta)$ are identical in their common domain of analyticity. Let $\sigma > \max(0, \sigma_n, \sigma_{n+1})$ and form the function

$$\Delta_n(\zeta) = (1 + \zeta)^{-\sigma-1}[W_n(\zeta) - W_{n+1}(\zeta)]. \tag{3.3.10}$$

The preceding discussion shows that $\Delta_n(\zeta)$ has the following properties:

(i) $\Delta_n(\zeta)$ is holomorphic for $\xi \geqq \sigma$.

(ii) $\| \Delta_n(\sigma + re^{i\theta}) \| \leqq M_n \exp\left[\dfrac{2r}{\alpha_n}\, l(\theta)\right]$.

(iii) $\Delta_n(k\alpha_n) = 0$ for all integers $k > \sigma/\alpha_n$.

(iv) $\dfrac{1}{\delta}\left[\dfrac{\pi}{2} - l\left(\dfrac{\pi}{2} - \delta\right)\right] \to +\infty$ as $\delta \to 0$.

These properties together imply that $\Delta_n(\zeta)$ is identically zero.[11] Hence

$$W_n(\zeta) \equiv W_{n+1}(\zeta), \qquad \xi > \max(\sigma_n, \sigma_{n+1}).$$

---

[11] See F. Carlson [1], theorem E on p. 60. For the reader's convenience the special case of this theorem which is needed here is reproduced in the Appendix of the present paper. Carlson's theorem applies directly to the functions $L[\Delta_n(\zeta)x]$, $x \in E$, $L \in \bar{E}$, and shows that they are all identically zero, whence it follows that $\Delta_n(\zeta) \equiv 0$.

It follows that all functions $W_n(\zeta)$ are identical, $n \geqq n_0$, and define the same analytic function $W(\zeta)$ which is represented by the convergent binomial series (3.3.5) for $h = \alpha_n$, $\xi > \sigma_n$.

Theorem 8 shows that $\alpha_n\tau(\alpha_n)$ tends to a finite non-negative limit, c say, as $n \to \infty$. Since $\sigma_n = \alpha_n[\tau(\alpha_n) - 1]$ we conclude that $W(\zeta)$ is holomorphic for $\xi > c$. It is bounded for $\xi \geqq c + \delta$, $\delta > 0$, on account of (3.3.7), since we can choose n so large that $\sigma_n < c + \delta$.

Formula (3.3.8) says that $W_n(\xi) = T_\xi$ for $\xi = k\alpha_n = k2^{-n}a_2$ provided these values exceed σ_n. It follows that $W(\xi) = T_\xi$ for every $\xi > c$ which is of the form $k2^{-n}a_e$, where k and n are arbitrary positive integers. These points are everywhere dense on (c, ∞). But T_ξ is by assumption (T_4) weakly continuous on the right for $\xi = \alpha_0$, hence also for every $\xi > \alpha_0$. From this we conclude that $L[W(\xi)x] \equiv L[T_\xi x]$ for $\xi > c$ and for every $x \in E$, $L \in \bar{E}$, whence it follows that $W(\xi) = T_\xi$, $\xi > c$. This completes the proof of the theorem.

Theorems 6 and 9 are not quite comparable unless $c = 0$. This will certainly be the case if

$$\alpha_n\tau(\alpha_n) \to 0 \quad \text{as} \quad n \to \infty. \tag{3.3.11}$$

It does not seem obvious that the domain of existence of $W(\zeta)$ as a holomorphic operator function in $\mathfrak{S}$ has to extend all the way to the imaginary axis, i.e., that c must be zero.

3.4. Spectral Properties of Semi-Groups, II

In the present paragraph we continue the discussion of the spectral properties of semi-groups which was started in §3.2. We assume now that (T_1) to (T_6) are satisfied, so that T_ξ has an analytic extension $W(\xi + i\eta)$.

It follows that all the results of §3.1 apply provided we replace the half-plane $\xi > 0$ by $\xi > c$. We see consequently that there exists an ω_1 such that for $0 < \omega < \omega_1$, the resolvent $R(\lambda, \omega)$ is holomorphic for $|\lambda - 1| > 1$ and of finite order $\rho(\omega)$ on the circle. This implies that $R(\lambda, \omega)$ is of finite order in the sector $V_{\pi/2} : 0 < |\lambda| \leqq 1$, $\pi/2 \leqq \arg\lambda \leqq 3\pi/2$ and this order cannot exceed $2\rho(\omega)$ as is seen by comparing the distances $|\lambda|$ and $|\lambda - 1| - 1$ from λ to the origin and to the circle respectively. Using the argument employed in the proof of Theorem 8, we easily verify that $R(\lambda, \omega)$ is of finite order in the sector $V_{\pi/4} : 0 < |\lambda| \leqq 1$, $\pi/4 \leqq \arg\lambda \leqq 7\pi/4$ for $\omega < \frac{1}{2}\omega_1$ provided $R(\lambda, \omega)$ is of sub-exponential order in this sector. The same type of argument applies to any other sector V_ψ with $0 < \psi < \pi/2$.

If $\tau(\omega)$ is the order of $R(\lambda, \omega)$ in some sector V_ψ it seems likely that $\omega\tau(\omega)$ is an increasing function of ω, but I can not prove this surmise.

Using Theorem 3 we see that

$$W(\zeta) = \sum_{n=0}^{\infty} (-1)^n [I - W(\omega)]^n \binom{\zeta/\omega}{n} \tag{3.4.1}$$

for $0 < \omega < \omega_1$, $\xi > \sigma_0(\omega)$. Here $\sigma_0(\omega) \leqq \omega[\tau(\omega) - 1]$ and decreases to c as $\omega \to 0$. We have evidently also

$$W(\zeta) = \frac{1}{2\pi i} \int_C R(\lambda, \omega)\lambda^{\zeta/\omega}\, d\lambda \tag{3.4.2}$$

for the same values of the variables. Here we can let C be any contour surrounding $S(\omega)$ once in the positive sense which starts and ends at the origin. Taking C to be the circle $|\lambda - r_\omega| = r_\omega$ we get the following result.

THEOREM 10. *There exists an $\Omega \geqq \omega_1$ such that for $0 < \omega < \Omega$ we can find circles $|\lambda - r_\omega| = r_\omega$ containing $S(\omega)$. For any such r_ω*

$$W(\zeta) = [r_\omega]^{\zeta/\omega} \sum_{n=0}^{\infty} (-1)^n [I - W(\omega)/r_\omega]^n \binom{\zeta/\omega}{n}, \tag{3.4.3}$$

where the series converges in the uniform sense at least for $\xi > \omega[\tau(\omega) - 1]$. If $r(\omega) = \inf r_\omega$ for fixed ω, then $(1/\omega) \log r(\omega)$ is an increasing function of ω.

PROOF. If r_ω is chosen so that the circle in question contains $S(\omega)$ in its interior except for the point $\lambda = 0$ we have

$$R(\lambda, \omega) = -\sum_{n=0}^{\infty} (r_\omega - \lambda)^{-n-1} [r_\omega I - W(\omega)]^n \tag{3.4.4}$$

for $|\lambda - r_\omega| > r_\omega$. If we suppose in addition that there exist a sector V_ψ with $\psi < \pi/2$ in which $R(\lambda, \omega)$ is of finite order $\tau(\omega)$, then we can substitute the series (3.4.4) in formula (3.4.2) and integrate term-wise. The resulting expansion is (3.4.3). The convergence is settled as in the proof of Theorem 9.

Now take a fixed value of ω, $\omega = \alpha$ say, and the corresponding series (3.4.4) with $r_\alpha > r(\alpha)$. Thus

$$W(\zeta)\, [r_\alpha]^{-\zeta/\alpha} = \sum_{n=0}^{\infty} (-1)^n [I - W(\alpha)/r_\alpha]^n \binom{\zeta/\alpha}{n}. \tag{3.4.5}$$

By Theorem 3 the same function also admits of an expansion in a binomial series in terms of the variable ζ/ω when $\omega < \alpha$, the abscissa of convergence of which is not greater than that of (3.4.5). This new series is

$$W(\zeta)\, [r_\alpha]^{-\zeta/\alpha} = \sum_{n=0}^{\infty} (-1)^n [I - W(\omega) r_\alpha^{-\omega/\alpha}]^n \binom{\zeta/\omega}{n}.$$

This is evidently an expansion of type (3.4.3), i.e. $r(\omega) \leqq r_\alpha^{\omega/\alpha}$. This holds for every $r_\alpha > r(\alpha)$, hence $r(\omega) \leqq [r(\alpha)]^{\omega/\alpha}$ or $(1/\omega) \log r(\omega)$ is an increasing function of ω. This completes the proof of the theorem.

The quantity Ω introduced in Theorem 10 may be finite or infinite. In the latter case we have the following result.

THEOREM 11. *If $\Omega = \infty$, i.e., if $S(\omega)$ is located in the right half-plane for every $\omega > 0$, then $S(\omega)$ is a real non-negative point set for every $\omega > 0$.*

PROOF. In order to see this, it is enough to refer to formula (3.2.4) according to which $[S(\omega)]^n \subset S(n\omega)$. If a non-real value of λ should belong to some $S(\omega)$,

then we can choose n so large that the n^{th} power of λ lies in the left half-plane. Then $\lambda^n \in S(n\omega)$ or $S(n\omega)$ has a point in the left half-plane against our assumption.

Suppose now that Ω is finite. We understand here by Ω the least upper bound of the values of ω for which $S(\omega)$ can be enclosed in a circle $|\lambda - r_\omega| = r_\omega$ and $R(\lambda, \omega)$ is of finite order in some sector V_ψ where $\psi = \psi(\omega) < \pi/2$. Then there cannot exist an a such that $\Omega < a < 2\Omega$ for which $S(a)$ can be enclosed in a circle $|\lambda - r_a| = r_a$ and $R(\lambda, \omega)$ is of finite order in some sector V_ψ for $\omega = a$ and of sub-exponential order for $\omega < a$. Indeed, Theorem 8 applies for such an a, and the argument used in proving Theorems 9 and 10 can be employed again to show that $R(\lambda, \omega)$ is holomorphic outside of the circle $|\lambda - r_a| = r_a$ and of finite order in V_ψ for $0 < \omega < a$. This contradicts the definition of Ω.

This does not exclude the possibility that $S(\omega)$ can be enclosed in a circle $|\lambda - r_\omega| = r_\omega$ and that $R(\lambda, \omega)$ is of finite order in V_ψ for some $\omega > 2\Omega$. The following example will make the situation clear. Let $\mathfrak{S} : \{W(\zeta)\}$ be a semi-group, holomorphic in $\xi > 0$, such that for each $\xi > 0$ the spectrum of $W(\xi)$ consists of the line-segment $(0, 1)$. We shall encounter such semi-groups in the next chapter. Then form $\mathfrak{S}_1 : \{U(\zeta)\}$ where $U(\zeta) = e^{i\zeta}W(\zeta)$. This is evidently also a semi-group holomorphic in $\xi > 0$. The spectrum of $U(\xi)$ consists of the line-segment joining the point $e^{i\xi}$ with the origin in the λ-plane. We can choose $W(\zeta)$ so that the order of $W(\xi)$ is one in any fixed sector V_ψ for every $\xi > 0$. For $\mathfrak{S}_1$ the constant Ω equals $\pi/2$, but for $(2k + 1)\pi/2 < \omega < (2k + 3)\pi/2$, $k = 1, 2, 3, \cdots$, $S(\omega)$ is located in the right half-plane and can be enclosed in a circle $|\lambda - r_\omega| = r_\omega$.

The last result of the present section is an extension of Theorem 7.

THEOREM 12. *Let $a \leqq \Omega$ and let $r(a)$ be determined as in Theorem* 10. *For $\omega < a$ the spectrum $S(\omega)$ of T_ω is enclosed in the region $S(\omega, a)$:*

$$|\lambda| \leqq \left\{2\, r(a) \cos \frac{a}{\omega}\theta\right\}^{\omega/a}, \qquad |\theta| \leqq \frac{\omega\pi}{2a}. \tag{3.4.6}$$

PROOF. We shall show that $R(\lambda, \omega)$ is holomorphic outside of $S(\omega, a)$. Let m be the least integer greater than $c/\omega - \frac{1}{2}$. Then for values of λ to be specified later[12]

$$\begin{aligned}\sum_{m+1}^{\infty} T_{n\omega}\lambda^{-n} &= \int_\mu^\infty W(\omega t)\lambda^{-t}\, dt \\ &\quad - i\lambda^{-\mu}\int_0^\infty \{W(\omega\mu + i\omega u)\lambda^{-iu} - W(\omega\mu - i\omega u)\lambda^{iu}\}\,[1 + e^{2\pi u}]^{-1}\, du \\ &\equiv I_1 + I_2.\end{aligned}$$

Here $\mu = m + \frac{1}{2}$ and the first integral is to be taken along a ray $\arg(t - \mu) = \psi$ to be chosen later. Substituting the estimate of $\| W(\zeta) \|$ which is obtained

[12] See E. Lindelöf [1], p. 129 et seq.

from formula (3.4.5) by the usual evaluation of the binomial series, we get after simple calculations

$$\| I_2 \| \leqq M \, | \lambda |^{-\mu}, \, 0 < | \lambda | < \infty, \, | \arg \lambda | < \frac{3\pi}{2},$$

and I_2 has no other singularities than 0 and ∞ as long as the argument of λ is restricted in the manner indicated.

Thus the only singularities which are of interest to us arise from I_1. The norm of the integrand in I_1 is dominated by

$$\rho^{\sigma} \exp \left\{ \rho \left[\frac{\omega}{a} \, l(\psi) + \cos \psi \cdot \log Q + \theta \sin \psi \right] \right\},$$

where $Q = r_a^{\omega/a} r^{-1}$, $r = | \lambda |$, $\theta = \arg \lambda$, and $\sigma > \sigma_a + \frac{1}{2}$. If $\frac{\omega\pi}{2a} \leqq | \theta | \leqq \pi$, and $2Q^{a/\omega} \leqq 1$, any choice of ψ such that $\operatorname{sgn} \psi = - \operatorname{sgn} \theta$ will give a convergent integral. If $2Q^{a/\omega} > 1$, we take $\cos \psi = \frac{1}{2} Q^{-a/\omega}$, $\operatorname{sgn} \psi = -\operatorname{sgn} \theta$, and get the same result. Thus for $\frac{\omega\pi}{2a} \leqq | \theta | \leqq \pi$ we find that $\| I_1 \|$ is finite and bounded in the interior of the sector. In particular, I_1 is holomorphic in this sector.

Suppose now that $| \theta | < \frac{\omega\pi}{2a}$. The multiplier of ρ in the exponent is

$$\left(\frac{\omega}{a} \psi + \theta \right) \sin \psi + \log \left\{ \left[2 r_a \cos \frac{a}{\omega} \theta \right]^{\omega/a} r^{-1} \right\} \cos \psi.$$

Let us choose $\psi = - \frac{a}{\omega} \theta$. The exponent then becomes

$$\log (R/r) \cos \frac{a}{\omega} \theta,$$

where

$$R = \left[2 r_a \cos \frac{a}{\omega} \theta \right]^{\omega/a}.$$

The exponent is negative for $r > R$. If this condition is satisfied

$$\| I_1 \| \leqq \Gamma(\sigma + 1) \left[\log (R/r) \cos \frac{a}{\omega} \theta \right]^{-\sigma-1}.$$

It follows that $R(\lambda, \omega)$ is holomorphic outside of the curve $r = R$ for every permissible choice of r_a. Hence we can replace r_a by its greatest lower bound $r(a)$ and the theorem is proved.[13]

---

[13] This proof is based upon an argument given by F. Carlson [2], pp. 49–52.

3.5. Semi-Groups with Real Spectra

We shall now consider semi-groups such that each transformation T_α has a real spectrum. We suppose that conditions (T_1) to (T_4) are satisfied and replace (T_5) and (T_6) by the following assumptions.

(T_5') *The spectrum of T_α is real positive for every $\alpha > 0$.*

(T_6') *If δ denotes the distance of λ from $S(\alpha)$, then there exist constants $K(\alpha)$ and $\pi(\alpha)$ such that for every $\alpha > 0$*

$$\| R(\lambda, \alpha) \| \leqq K(\alpha)\delta^{-\pi(\alpha)}$$

It is clear that these assumptions imply (T_5) and (T_6) respectively so that Theorems 9, 10, and 12 hold. In particular, there exists an analytic extension $W(\zeta)$ of T_α defined for $\xi > c$ where

$$0 \leqq c \leqq \liminf_{\alpha \to 0} \alpha\pi(\alpha). \tag{3.5.1}$$

This function $W(\zeta)$ is representable by the various binomial series which we have encountered in §§3.3 and 3.4, but in the present case other representations are available, namely in terms of Laplace-Stieltjes integrals.

The spectrum of T_α is supposed to be real positive. We can always reduce the discussion to the case in which the spectrum is restricted to the interval $[0, 1]$. In order to do so, let us recall the quantity $r(\alpha) = \inf r_\alpha$ introduced in Theorem 10. Put

$$2r(\alpha) = M(\alpha), \qquad M(1) = M. \tag{3.5.2}$$

Then $M(\alpha)$ is the least upper bound of the spectral values of T_α. We shall show that

$$M(\alpha) \leqq M^\alpha.$$

Theorem 12 shows that this is true for $0 < \alpha \leqq 1$. Suppose that it has been shown to be true for $0 < \alpha \leqq a$. In §3.2 it was proved that if $\lambda_0 \in S(\alpha)$, then either $+\sqrt{\lambda_0}$ or $-\sqrt{\lambda_0}$ belongs to $S(\alpha/2)$. In the case of real positive spectra we must take the positive square root. Suppose $a < \alpha \leqq 2a$ and that $\lambda_0 > M^\alpha$. Then $\sqrt{\lambda_0} > M^{\alpha/2}$ and a fortiori, $M(\alpha/2) > M^{\alpha/2}$. This contradicts our assumption. Hence $M(\alpha) \leqq M^\alpha$ for all $\alpha > 0$.

Let us now define

$$U(\zeta) = M^{-\zeta}W(\zeta). \tag{3.5.3}$$

This is also an operator function having the semi-group property in the half-plane $\xi > c$. If $\lambda_0 \in S_W(\alpha)$ then $M^{-\alpha}\lambda_0 \in S_U(\alpha)$ and vice versa. It follows that $S_U(\alpha)$ is enclosed in the interval $[0, M(\alpha)M^{-\alpha}] \subset [0, 1]$. Hence the spectrum of $U(\alpha)$ always belongs to the interval $[0, 1]$.

We have now

$$U(\zeta) = \frac{1}{2\pi i}\int_C R_U(\lambda, \alpha)\lambda^{\zeta/\alpha}\, d\lambda, \tag{3.5.4}$$

where C is any rectifiable contour surrounding the interval [0, 1] once in the positive sense, beginning and ending at the origin, and $\alpha > 0$ is arbitrary. The integral has a sense at least for $\xi > \alpha[\pi(\alpha) - 1]$.

We take a particular contour C in order to estimate the norm of the integral operator. Let ρ be a small number to be disposed of later and let C consist of the two straight line segments joining the origin with the points $1 + \rho i$ and $1 - \rho i$ and the semi-circle of radius ρ joining the points and passing to the right of $+1$. The rectilinear paths give contributions to the norm of the form

$$K_1(\alpha)\,[1 + \rho^2]^{\xi/2\alpha} \exp\left[\frac{|\eta|}{\alpha} \arctan \rho\right] \{\xi - \alpha[\pi(\alpha) - 1]\}^{-1}\, \rho^{-\pi(\alpha)},$$

whereas the circular arc gives

$$K_2(\alpha)(1 + \rho)^{\xi/\alpha} \exp\left[\frac{|\eta|}{\alpha} \arctan \rho\right] \rho^{-\pi(\alpha)+1}.$$

If ξ is fixed and $> \alpha\,[\pi(\alpha) - 1]$ we can choose $\rho = \alpha/|\eta|$ and obtain

$$\| U(\xi + i\eta) \| \leqq K(\alpha, \xi)\, |\eta|^{\pi(\alpha)}, \tag{3.5.5}$$

whereas for $\zeta = c(\alpha) + re^{i\theta}$, $|\theta| < \pi/2$, we can take $\rho = \alpha/r$ and get

$$\| U(c(\alpha) + re^{i\theta}) \| \leqq K(\alpha, \theta) r^{\pi(\alpha)-1}, \tag{3.5.6}$$

where $c(\alpha)$ is any quantity greater than $\alpha[\pi(\alpha) - 1]$. These are the required estimates.

Let us now introduce the Lindelöf mu-function.

$$\mu(\xi) = \limsup_{|\eta| \to \infty} \frac{\log \| U(\xi + i\eta) \|}{\log |\eta|}, \tag{3.5.7}$$

which is defined for $\xi > c$. It follows from (3.5.5) that $\mu(\xi) \leqq \pi(\alpha)$ if $\xi > \alpha\,[\pi(\alpha) - 1]$. It is known that $\mu(\xi)$ is a convex, monotone decreasing function, continuous for $\xi > c$. $W(\zeta)$ and $U(\zeta)$ have the same mu-functions.

Let $\xi_0 > c$, $\beta > \mu(\xi_0) + 1$, and form

$$A_\beta(t) = \frac{1}{2\pi i} \int_{\xi_0 - i\infty}^{\xi_0 + i\infty} e^{t\zeta} U(\zeta) \zeta^{-1-\beta}\, d\zeta. \tag{3.5.8}$$

This defines an operator function depending upon two parameters t and β. It is easy to see that $A_\beta(t) = 0$ for $t \leqq 0$ and

$$\| A_\beta(t) \| \leqq K_1 e^{t\xi_0}, \int_0^t \left\| \frac{d}{du} A_\beta(u) \right\| du \leqq K_2 e^{t\xi_0}. \tag{3.5.9}$$

Thus $A_\beta(t)$ is absolutely continuous even in the uniform sense.

Next form the integral

$$\int_0^\infty e^{-\zeta t}\, dA_\beta(t) \equiv \int_0^\infty e^{-\zeta t} \frac{d}{dt} A_\beta(t)\, dt. \tag{3.5.10}$$

which is actually of the Laplace type though we use the Laplace-Stieltjes notation to indicate the possibilities of generalizations. The estimates (3.5.9) suffice to show the absolute convergence of this integral for $\xi > \xi_0$ in the sense that

$$\int_0^\infty e^{-\xi t} \left\| \frac{d}{dt} A_\beta(t) \right\| dt < \infty. \tag{3.5.11}$$

The argument which is familiar in the theory of ordinary Laplace-Stieltjes integrals can be used to show that the value of the integral in (3.5.10) is $\zeta^{-\beta}U(\zeta)$. Thus

$$U(\zeta) = \zeta^\beta \int_0^\infty e^{-\zeta t}\, dA_\beta(t) \tag{3.5.12}$$

for $\xi > \xi_0$, $\beta > \mu(\xi_0) + 1$.

The last condition is unnecessarily restrictive. $A_\beta(t)$ has a sense at least for $\beta > \mu(\xi_0)$, but may cease to be absolutely continuous for $\beta < \mu(\xi_0) - 1$ and may even lose the property of being of bounded variation.[14] But formula (3.5.12) is still valid if the integral is suitably interpreted. It suffices to take

$$\int_0^\infty e^{-\zeta t}\, dA_\beta(t) = \zeta \int_0^\infty e^{-\zeta t} A_\beta(t)\, dt. \tag{3.5.13}$$

The integral on the right converges absolutely for $\xi > \xi_0$ in the sense that it converges if we replace $e^{-\zeta t}$ by its absolute value and $A_\beta(t)$ by its norm.

The dependence of $A_\beta(t)$ upon the parameter β is made evident by the formula

$$A_{\alpha+h}(t) = \frac{1}{\Gamma(h)} \int_0^t (t-s)^{h-1} A_\alpha(s)\, ds.$$

We have thus proved

THEOREM 13. *Let $\mathfrak{S}: \{T_\alpha\}$ satisfy postulates (T_1) to (T_4), (T_5') and (T_6'). Then there exists an analytic extension $W(\zeta)$ of T_α, holomorphic and having the semi-group property for $\xi > c$, where c satisfies* (3.5.1). *The Lindelöf mu-function $\mu(\xi)$ of $W(\zeta)$ exists and is bounded for $\xi \geqq c + \delta$. For $\xi > \xi_0 > c$ there exist representations of the form*

$$W(\zeta) = M^\zeta \zeta^{1+\beta} \int_0^\infty e^{-\zeta t} A_\beta(t)\, dt, \tag{3.5.14}$$

where $A_\beta(t)$ is given by (3.5.8) *at least for $\beta > \mu(\xi_0)$. If the integrand is replaced by its norm, the resulting integral converges for $\xi > \xi_0$.*

The most interesting special case is that in which $\pi(\alpha)$ is bounded $\leqq \pi_0$ for $\alpha > 0$. We have then $c = 0$, $\mu(\xi) \leqq \pi_0$ for $\xi > 0$, and (3.5.14) is valid at least for $\beta > \pi_0$, the integral being convergent for $\xi > 0$.

---

[14] Cf. E. Hille & J. D. Tamarkin [2] for the corresponding phenomenon in connection with ordinary Laplace-Stieltjes integrals.

3.6. The Embedding Theorem

We shall take up the question of constructing one-parameter families of transformations having the semi-group property in a half-plane. We shall show that under suitable restrictions on the resolvent of a transformation T it is possible to embed T in such a family of transformations by a suitable process of interpolation performed on the positive integral powers of T.

THEOREM 14. *Let $T \in \mathfrak{S}$ and let $R(\lambda)$, the resolvent of T, be holomorphic outside of the sector $V_s : 0 \leqq |\lambda| \leqq K, \psi_1 \leqq \arg\lambda \leqq \psi_2, \psi_2 - \psi_1 < \pi$, and let $\| R(\lambda) \| \leqq M |\lambda|^{-\tau}$ for $|\lambda| < K$ in the complementary sector. Let*

$$W(\zeta) = \frac{1}{2\pi i}\int_C R(\lambda)\lambda^{\zeta}\, d\lambda, \tag{3.6.1}$$

where C surrounds V_s once in the positive sense, beginning and ending at the origin. Then $W(\zeta)$ is holomorphic and has the semi-group property at least for $\xi > \tau - 1$ and $W(n) = T^n$ for $n > \tau - 1$. $W(\zeta)$ generates a semi-group if $\tau = 1$ and a group if $\tau < 1$.

PROOF. Let $\psi_0 = \frac{1}{2}(\psi_1 + \psi_2)$ and suppose $-\pi \leqq \psi_0 < +\pi$. On C we suppose $\psi_0 - \pi/2 \leqq \arg\lambda \leqq \psi_0 + \pi/2$ and take $\lambda^{\zeta} = \exp[\zeta(\log|\lambda| + i\arg\lambda)]$. The integral in (3.6.1) is easily seen to exist as a holomorphic function for $\xi > \tau - 1$ and $W(n) = T^n$ in the same half-plane by Cauchy's theorem. In order to see the semi-group property we had better expand $W(\zeta)$ in a binomial series. Let $a = \rho_0 e^{i\psi_0}$ where $\rho_0 > K/2$. A simple calculation shows that

$$W(\zeta) = a^{\zeta}\sum_{n=0}^{\infty}(-1)^n(I - T/a)^n\binom{\zeta}{n}, \tag{3.6.2}$$

where the sum of the norms converges for $\xi > \tau - 1$. Now form $W(\zeta_1)[W(\zeta_2)]$ with $\Re(\zeta_\nu) > \tau - 1$ and substitute the series. Since the powers of $(I - T/a)$ commute with each other and with the scalar factors, we obtain the double series

$$a^{\zeta_1+\zeta_2}\sum_{j=0}^{\infty}\sum_{k=0}^{\infty}(-1)^{j+k}(I - T/a)^{j+k}\binom{\zeta_1}{j}\binom{\zeta_2}{k}$$

which converges if every term is replaced by its norm. Hence we can collect terms freely and obtain the required result $W(\zeta_1 + \zeta_2)$.

$W(\zeta)$ is holomorphic in the right half-plane if $\tau = 1$ and hence generates a semi-group in this case. If $\tau < 1$, we find that $W(\zeta)$ generates a group instead. Indeed, $W(\zeta)$ is then holomorphic in a strip in the left half-plane and can consequently be continued analytically in the whole plane by Theorem 5. This case arises in particular if $\lambda = 0$ does not belong to the spectrum of T.

A similar remark applies to Theorems 9 and 13. If for any choice of $\alpha > 0$ the quantity $\tau(\alpha)$ or $\pi(\alpha)$ is less than one the corresponding function $W(\zeta)$ generates a group.

The function $W(\zeta)$ defined by (3.6.1) is not the only solution of the interpolation problem. If k is any integer, we can clearly multiply $W(\zeta)$ by $e^{2k\pi i\zeta}$ and still get a solution. It is not unlikely that these are the only measurable solu-

tions, but I cannot prove that this is the case. At any rate the solution $W(\zeta)$ has an extremal property. If $U(\zeta)$ is any other solution of the interpolation problem which is holomorphic in a right half-plane, then $U(\zeta) - W(\zeta)$ vanishes at all the integers in their common half-plane of existence. If this difference $D(\zeta)$ is not to vanish identically, its norm must satisfy the following inequality[15]

$$\limsup_{|\eta|\to\infty} \frac{1}{|\eta|} \log \| D(\xi + i\eta) \| \geqq \pi.$$

On the other hand, direct estimates of the integral (3.6.1) shows that

$$\limsup_{|\eta|\to\infty} \frac{1}{|\eta|} \log \| W(\xi + i\eta) \| \leqq \max (|\psi_1|, |\psi_2|). \tag{3.6.3}$$

Hence if $-\pi < \psi_1 < \psi_2 < \pi$, we can claim that $W(\zeta)$ *gives the solution of the interpolation problem with the minimum rate of growth of the norm on vertical lines.* These limits are precise. For a suitable choice of a transformation T with a real negative spectrum, say on the interval $[-1, 0]$, so that $\psi_1 = \psi_2 = \pi$, the two transformations

$$e^{\pm\pi i\zeta} \sum_{n=0}^{\infty} (-1)^n (I + T)^n \binom{\zeta}{n}$$

may have the same rate of growth.

4. Special Instances of the General Theory

4.1. Self-Adjoint Semi-Groups in Hilbert Space

Suppose that $E = \mathfrak{H}$, the Hilbert space, and let $\mathfrak{S} : \{T_\alpha\}$ be a semi-group of transformations on $\mathfrak{H}$ to $\mathfrak{H}$ defined for $\alpha \geqq 0$. In addition we suppose that

(T_7) *T_α is self-adjoint for every $\alpha > 0$.*

This means that

$$(T_\alpha x, y) = (x, T_\alpha y), \qquad x, y \in \mathfrak{H}, \tag{4.1.1}$$

where (u, v) is the fundamental bilinear functional which defines the metric of the space. As the first consequence of this relation we get

$$\begin{aligned} (T_\alpha x, x) &= (T_{\alpha/2} T_{\alpha/2} x, x) \\ &= (T_{\alpha/2} x, T_{\alpha/2} x) = \| T_{\alpha/2} x \|^2 \geqq 0, \end{aligned} \tag{4.1.2}$$

i.e., T_α is *positive definite* for every $\alpha > 0$.

Since $|(T_\alpha x, T_\beta x)| \leqq \| T_\alpha x \| \, \| T_\beta x \|$, we get

$$(T_{\alpha+\beta} x, x)^2 \leqq (T_{2\alpha} x, x)(T_{2\beta} x, x). \tag{4.1.3}$$

Hence putting

$$\log (T_\alpha x, x) = \varphi(\alpha; x), \tag{4.1.4}$$

[15] Cf. F. Carlson [1], Theorem C, p. 58.

we find that

$$\varphi(\tfrac{1}{2}(\alpha+\beta);x) \leqq \tfrac{1}{2}[\varphi(\alpha;x)+\varphi(\beta;x)]. \tag{4.1.5}$$

In other words, $\varphi(\alpha;x)$ is convex with respect to α.[16]

In postulate (T_3) we assume that $\| T_\alpha \|$ is measurable for $\alpha > 0$. In the present case it is more convenient to replace this postulate by either of the following assumptions.

(T_3') $\| T_\alpha x \|$ *is measurable for every* $x \in \mathfrak{H}$.

(T_3'') *The limits of indetermination of* $\| T_\alpha x \|$ *as* $\alpha \to 0$ *are finite for every* $x \in \mathfrak{H}$.

Assumption (T_3') implies that $\varphi(\alpha;x)$ is measurable for $\alpha > 0$. A convex function is continuous in the interior of its interval of definition if it is measurable. Hence $\varphi(\alpha;x)$ is continuous for $\alpha > 0$, x fixed, and the same is true of $(T_\alpha x, x)$ and $\| T_\alpha x \|$. From this we conclude that T_α is weakly continuous, i.e., (T_3') implies (T_4). That (T_3') implies (T_3) follows from the separability of the space.

If we use (T_3'') instead of (T_3') we notice that $\varphi(\alpha;x)$ is a convex function bounded above at the origin and hence bounded above in every finite interval by the semi-group property. Such a function is continuous and we can proceed as above.

From the fact that T_α is a positive definite transformation follows that its spectrum is real and non-negative. If M is defined by formula (3.5.2), we know that $S(\alpha)$ is located in the interval $[0, M^\alpha]$. Thus postulates (T_5) and (T_5') are satisfied.

Put $\lambda = \mu + i\nu$. Then

$$\| (T_\alpha - \lambda I)x \|^2 = (T_{2\alpha}x, x) - 2\mu(T_\alpha x, x) + (\mu^2+\nu^2)(x,x). \tag{4.1.6}$$

Disregarding for a moment the term in ν^2, what remains is a quadratic polynomial in μ which is never negative by virtue of (4.1.3). We have obviously

$$\| T_\alpha - \lambda I \| \geqq |\lambda|, \qquad \mu \leqq 0, \tag{4.1.7}$$

$$\| T_\alpha - \lambda I \| \geqq |\nu|, \qquad \mu > 0, \tag{4.1.8}$$

---

[16] The convexity of $\log (T_\alpha x, x)$ was proved by B. v. Sz. Nagy [1] in the case of groups of self-adjoint transformations in Hilbert space and he based one of his proofs of the representation theorem for such groups upon this property. Dr. B. v. Sz. Nagy kindly called my attention to his paper in a letter of April 7, 1938, and indicated how his methods could be used to prove a more general representation theorem for semi-groups than the one stated and proved in outline in my note [3]. See Theorem 15 below. My original method had no points of contact with those of B. v. Sz. Nagy. In the present paragraph I develop both methods and a mixed one having elements in common with both of them. Postulate (T_3'') below was used by B. v. Sz. Nagy, but that Theorem 15 is valid under the assumptions (T_1), (T_2), (T_3''), and (T_7) was pointed out to the author by J. von Neumann in a letter of February 6, 1938, i.e., before the author was aware of B. v. Sz. Nagy's investigation.

but if $\mu > M^\alpha$ we can replace the second inequality by

$$\| T_\alpha - \lambda I \| \geqq | \lambda - M^\alpha | , \tag{4.1.9}$$

since

$$(T_\alpha x, x) = \| T_{\alpha/2} x \|^2 \leqq M^\alpha (x, x).$$

These inequalities taken together show that postulate (T_6') is satisfied and that

$$\| R(\lambda, \alpha) \| \leqq 1/\delta, \tag{4.1.10}$$

where δ is the distance from λ to $S(\alpha)$.

We have consequently shown that postulates (T_1), (T_2), (T_3') or (T_3''), and (T_7) are sufficient for the validity of Theorem 13. Thus T_α admits of an analytic extension $W(\zeta)$ in the right half-plane and

$$W(\zeta) = M^\zeta \zeta^{1+\beta} \int_0^\infty e^{-\zeta t} A_\beta(t)\, dt \tag{4.1.11}$$

for any $\beta > 1$.

Actually the representation is valid for $\beta \geqq 0$. It does not seem possible to prove this by a further refinement of the estimates of the norm of $W(\zeta)$. There are, however, various possibilities of proving (4.1.11) directly with little or no reference to the general theory developed in this paper. Thanks to the special properties of self-adjoint transformations in Hilbert space, special methods lead to a sharper result.

One possibility is offered by the canonical representation of self-adjoint transformations in Hilbert space, i.e.,

$$(Tx, y) = \int_{-\infty}^{\infty} s\, d(F(s)x, y),$$

where $F(s)$ is the resolution of the identity corresponding to T. Since T is positive definite and bounded and its spectrum belongs to the interval $[0, M]$ this formula reduces to

$$(Tx, y) = \int_0^M s\, d(F(s)x, y). \tag{4.1.12}$$

Let us now form

$$U(\zeta) = \int_0^M s^\zeta\, dF(s), \qquad \Re(\zeta) \geqq 0. \tag{4.1.13}$$

Using the properties of $F(s)$ we see that $U(\zeta)$ defines a semi-group in the right half-plane where it is holomorphic and satisfies the inequality

$$\| U(\xi + i\eta) \| \leqq M^\xi.$$

Further $U(n) = T^n$ for $n = 0, 1, 2, \cdots$. On the other hand, the function $W(\zeta)$, which we have constructed, also interpolates the sequence $\{T^n\}$ and has

the smallest rate of growth for its norm on vertical lines (cf. the remarks at the end of §3.6). It follows that $U(\zeta) = W(\zeta)$.

Hence

$$W(\zeta) = M^{\zeta} \int_0^{\infty} e^{-\zeta u}\, dE(u), \tag{4.1.14}$$

where $E(u) = F(Me^{-u})$. It is clear that $E(u)$ is also a resolution of the identity; the corresponding transformation can be interpreted as a determination of $-\log T$. As a rule this is not a bounded transformation. Its domain is that sub-set of $\mathfrak{H}$ for which

$$\int_0^{\infty} u^2\, d_u(E(u)x, x) < \infty.$$

We note that $-\log T$ is positive definite since $E(u) = 0$ for $u \leqq 0$.

We can arrive at the same result without any reference to the preceding general theory by observing that $U(\xi) = T_{\xi}$ for dyadic-rational values of ξ on account of the semi-group property and the positive definiteness of the transformations. Since $U(\xi)$ is weakly continuous by definition for $\xi > 0$ and T_{ξ} has been shown to have this property if we assume (T_3') or (T_3''), we see that $T_{\xi} = U(\xi)$ for all ξ. This is the argument of B. v. Sz. Nagy referred to in footnote 16.

Another possibility is offered by the properties of monotony which characterize $(T_{\alpha}x, x)$ as function of α for fixed x. Let us put

$$T_{\alpha} = M^{\alpha} V_{\alpha}. \tag{4.1.15}$$

Then $\mathfrak{S}_1 : \{V_{\alpha}\}$ is a semi-group of self-adjoint transformations having their spectra in the interval $[0, 1]$. Let $h > 0$, $\alpha \geqq 0$, and form

$$\Delta_h^n V_{\alpha} \equiv \sum_{k=0}^{n} (-1)^k \binom{n}{k} T_{\alpha+kh} = (I - V_h)^n V_{\alpha}. \tag{4.1.16}$$

This is a self-adjoint, positive definite transformation. In order to prove this, we notice that $(1 - t)^n$ is positive for $0 < t < 1$, whence it follows that $(I - V_h)^n$, which is obviously self-adjoint, is also positive definite.[17] Since V_{α} has the same properties and commutes with the powers of $I - V_h$ we conclude that (4.1.16) defines a self-adjoint, positive definite transformation. Hence

$$\left(\Delta_h^n V_{\alpha} x, x\right) \geqq 0$$

for $n = 0, 1, 2, \cdots$, and every $h > 0$, $\alpha \geqq 0$. In other words, $(V_{\alpha}x, x)$ is a completely monotone function of α for fixed x. It follows from the investigations of F. Hausdorff [1], S. Bernstein [1], and D. V. Widder [1], that

$$(V_{\alpha}x, x) = \int_0^{\infty} e^{-\alpha u}\, d_u E(u; x), \tag{4.1.17}$$

[17] See F. Riesz [11, pp. 31–33.

where $E(u; x)$ is a never decreasing function of u, $E(0; x) = 0$ and $E(u; x) \to (x, x)$ as $u \to \infty$.

The simplest way of completing the proof is to use formula (4.1.2) which tells us that

$$(V_n x, x) = \int_0^\infty e^{-nu}\, d_u(E(u)x, x)$$

for every positive integer n. Comparing with (4.1.17) and using the theorem of Lerch for Laplace-Stieltjes integrals, we see that $E(u; x) = (E(u)x, x)$ at all points of continuity. Thus we are again led to formula (4.1.14).

It is also possible to prove that $E(u; x) = (E(u)x, x)$, where $E(u)$ is a resolution of the identity, without using the canonical representation of T. The proof is too long to give in detail, but some indications of the various steps are not quite devoid of interest. One obtains from (4.1.17) the fact that

$$(V_\alpha x, y) = \int_0^\infty e^{-\alpha u}\, d_u E(u; x, y).$$

Here $E(u; x, y)$ is a bilinear functional in x and y which is of uniformly bounded variation with respect to u on $[0, \infty]$ if $\| x \| = \| y \| = 1$. Further $E(u; x, x) = E(u; x)$. With the aid of these facts one proves that $E(u; x, y) = (R(u)x, y)$, where $R(u)$ is a self-adjoint, positive definite transformation such that $0 \leqq (R(u)x, x) \leqq (x, x)$. Next one proves that $R(u)$ commutes with V_α, for instance, with the aid of the inversion formula of Laplace integrals and the semi-group property of V_α. This fact and the relation

$$(V_{\alpha+\beta} x, y) = \int_0^\infty e^{-\beta u}\, d_u(R(u)V_\alpha x, y)$$

give the identity

$$\int_0^u e^{-\alpha t}\, d_t(R(t)x, y) = \int_0^\infty e^{-\alpha t}\, d_t(R(u)R(t)x, y),$$

from which we conclude that $R(u)R(t) = R(s)$, where s is the smaller of the numbers t and u. These relations show that $R(u)$ is a resolution of the identity, provided we normalize in the conventional manner at the points of discontinuity.

We have consequently proved in three different ways the following result.

THEOREM 15. *Let $\mathfrak{S} : \{T_\alpha\}$ be a semi-group of transformations on $\mathfrak{H}$ to $\mathfrak{H}$ satisfying postulates* (T_1), (T_2), (T_3') *or* (T_3''), *and* (T_7). *Then T_α admits of an analytic extension $W(\zeta)$ in the right half-plane, and*

$$W(\zeta) = M^\zeta \int_0^\infty e^{-\zeta u}\, dE(u),$$

where $E(u)$ is the resolution of the identity of a positive definite self-adjoint transformation which may be regarded as a determination of $-\log T$.[18]

4.2. Special Examples

To illustrate Theorem 15 we shall discuss the transformations of Gauss-Weierstrass and Poisson defined by formulas (1.1.2) and (1.1.3) of the Introduction.

Let $x(t) \in L_2(-\infty, \infty)$ and form

$$W_\zeta x(t) = (\pi\zeta)^{-\frac{1}{2}} \int_{-\infty}^{\infty} \exp[-u^2/\zeta] x(t+u)\, du, \qquad \Re(\zeta) > 0. \tag{4.2.1}$$

It is a simple matter to show that $W_\zeta x(t) \in L_2(-\infty, \infty)$ and that $\mathfrak{S} = \{W_\zeta\}$ is a semi-group in the right half-plane. If $\zeta = \xi$ is real, it is easily seen that the transformation is self-adjoint and hence positive definite.

In order to study the spectrum of W_ζ and to determine its canonical representation as a Laplace-Stieltjes integral, we shall need a couple of other functional transformations of $x(t)$. We recall that the Fourier transform $X(t)$ of $x(t)$ is defined as

$$X(t) = \text{l.i.m.}\ (2\pi)^{-\frac{1}{2}} \int_{-a}^{a} e^{-itu} x(u)\, du,$$

and that $X(t) \in L_2$ and $\| X(t) \| = \| x(t) \|$. Further we need the Dirichlet transform $D_\alpha x(t)$ defined by

$$D_\alpha x(t) = \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{\sin \alpha u}{u} x(t+u)\, du, \qquad \alpha > 0. \tag{4.2.2}$$

This function has the properties[19] $D_\alpha x(t) \in L_2$, $\| D_\alpha x(t) \| \leqq \| x(t) \|$, and

$$(D_\alpha x(t), x(t)) = \int_{-\alpha}^{\alpha} |X(t)|^2\, dt. \tag{4.2.3}$$

We have then

$$\begin{aligned}(W_\zeta x, x) &= (\pi\zeta)^{-\frac{1}{2}} \int_{-\infty}^{\infty} \exp[-u^2/\zeta]\, du \int_{-\infty}^{\infty} x(t+u)\overline{x(t)}\, dt \\ &= \int_{-\infty}^{\infty} \exp[-\zeta s^2/4]\, |X(s)|^2\, ds,\end{aligned}$$

where we have used a classical relation in the theory of Fourier transforms. This formula shows that for ξ real we have

$$0 \leqq (W_\xi x, x) \leqq (x, x).$$

---

[18] In my note [3] a somewhat weaker theorem was stated and proved with the aid of the last of the three methods presented above. The theorem involved a boundedness assumption, viz. $\| I - T_\alpha \| \leqq 1$ for all α, which implies $M = 1$.

[19] For the properties of the Dirichlet transform used here, see E. Hille [2], pp. 150–153.

where $E(u; x)$ is a never decreasing function of u, $E(0; x) = 0$ and $E(u; x) \to (x, x)$ as $u \to \infty$.

The simplest way of completing the proof is to use formula (4.1.2) which tells us that

$$(V_n x, x) = \int_0^\infty e^{-nu}\, d_u(E(u)x, x)$$

for every positive integer n. Comparing with (4.1.17) and using the theorem of Lerch for Laplace-Stieltjes integrals, we see that $E(u; x) = (E(u)x, x)$ at all points of continuity. Thus we are again led to formula (4.1.14).

It is also possible to prove that $E(u; x) = (E(u)x, x)$, where $E(u)$ is a resolution of the identity, without using the canonical representation of T. The proof is too long to give in detail, but some indications of the various steps are not quite devoid of interest. One obtains from (4.1.17) the fact that

$$(V_\alpha x, y) = \int_0^\infty e^{-\alpha u}\, d_u E(u; x, y).$$

Here $E(u; x, y)$ is a bilinear functional in x and y which is of uniformly bounded variation with respect to u on $[0, \infty]$ if $\| x \| = \| y \| = 1$. Further $E(u; x, x) = E(u; x)$. With the aid of these facts one proves that $E(u; x, y) = (R(u)x, y)$, where $R(u)$ is a self-adjoint, positive definite transformation such that $0 \leqq (R(u)x, x) \leqq (x, x)$. Next one proves that $R(u)$ commutes with V_α, for instance, with the aid of the inversion formula of Laplace integrals and the semi-group property of V_α. This fact and the relation

$$(V_{\alpha+\beta} x, y) = \int_0^\infty e^{-\beta u}\, d_u(R(u)V_\alpha x, y)$$

give the identity

$$\int_0^u e^{-\alpha t}\, d_t(R(t)x, y) = \int_0^\infty e^{-\alpha t}\, d_t(R(u)R(t)x, y),$$

from which we conclude that $R(u)R(t) = R(s)$, where s is the smaller of the numbers t and u. These relations show that $R(u)$ is a resolution of the identity, provided we normalize in the conventional manner at the points of discontinuity.

We have consequently proved in three different ways the following result.

THEOREM 15. *Let $\mathfrak{S} : \{T_\alpha\}$ be a semi-group of transformations on $\mathfrak{H}$ to $\mathfrak{H}$ satisfying postulates* (T_1), (T_2), (T_3') *or* (T_3''), *and* (T_7). *Then T_α admits of an analytic extension $W(\zeta)$ in the right half-plane, and*

$$W(\zeta) = M^\zeta \int_0^\infty e^{-\zeta u}\, dE(u),$$

where $E(u)$ is the resolution of the identity of a positive definite self-adjoint transformation which may be regarded as a determination of $-\log T$.[18]

4.2. Special Examples

To illustrate Theorem 15 we shall discuss the transformations of Gauss-Weierstrass and Poisson defined by formulas (1.1.2) and (1.1.3) of the Introduction.

Let $x(t) \in L_2(-\infty, \infty)$ and form

$$(4.2.1) \qquad W_\zeta x(t) = (\pi\zeta)^{-\frac{1}{2}} \int_{-\infty}^{\infty} \exp[-u^2/\zeta] x(t+u)\, du, \qquad \Re(\zeta) > 0.$$

It is a simple matter to show that $W_\zeta x(t) \in L_2(-\infty, \infty)$ and that $\mathfrak{S} = \{W_\zeta\}$ is a semi-group in the right half-plane. If $\zeta = \xi$ is real, it is easily seen that the transformation is self-adjoint and hence positive definite.

In order to study the spectrum of W_ζ and to determine its canonical representation as a Laplace-Stieltjes integral, we shall need a couple of other functional transformations of $x(t)$. We recall that the Fourier transform $X(t)$ of $x(t)$ is defined as

$$X(t) = \text{l.i.m.}\ (2\pi)^{-\frac{1}{2}} \int_{-a}^{a} e^{-itu} x(u)\, du,$$

and that $X(t) \in L_2$ and $\| X(t) \| = \| x(t) \|$. Further we need the Dirichlet transform $D_\alpha x(t)$ defined by

$$(4.2.2) \qquad D_\alpha x(t) = \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{\sin \alpha u}{u} x(t+u)\, du, \qquad \alpha > 0.$$

This function has the properties[19] $D_\alpha x(t) \in L_2$, $\| D_\alpha x(t) \| \leqq \| x(t) \|$, and

$$(4.2.3) \qquad (D_\alpha x(t), x(t)) = \int_{-\alpha}^{\alpha} |X(t)|^2\, dt.$$

We have then

$$(W_\zeta x, x) = (\pi\zeta)^{-\frac{1}{2}} \int_{-\infty}^{\infty} \exp[-u^2/\zeta]\, du \int_{-\infty}^{\infty} x(t+u)\overline{x(t)}\, dt$$

$$= \int_{-\infty}^{\infty} \exp[-\zeta s^2/4]\, |X(s)|^2\, ds,$$

where we have used a classical relation in the theory of Fourier transforms. This formula shows that for ξ real we have

$$0 \leqq (W_\xi x, x) \leqq (x, x).$$

---

[18] In my note [3] a somewhat weaker theorem was stated and proved with the aid of the last of the three methods presented above. The theorem involved a boundedness assumption, viz. $\| I - T_\alpha \| \leqq 1$ for all α, which implies $M = 1$.

[19] For the properties of the Dirichlet transform used here, see E. Hille [2], pp. 150–153.

Moreover, if we give any λ, $0 < \lambda < 1$, and any $\xi > 0$, we can find an $x(t)$ with $\| x(t) \| = 1$ such that $(W_\xi x(t), x(t)) = \lambda$. For this purpose it is enough to choose

$$x(t) = \frac{\sin bt}{\sqrt{(b\pi)t}}, \qquad (W_\xi x, x) = A^{-1} \int_0^A e^{-u^2} du, \quad A = \tfrac{1}{2} b \sqrt{\xi},$$

and to determine A so that the integral equals λA which is possible when $0 < \lambda < 1$. Every point of the interval [0, 1] belongs to the spectrum of W_ξ.

Using formula (4.2.3) we see that

$$(W_\xi x, x) = \int_0^\infty e^{-\xi s^2/4} d_s(D_s x, x)$$

or

$$(W_\xi x, x) = \int_0^\infty e^{-\xi u} d_u(D_{2\sqrt{u}} x, x). \tag{4.2.4}$$

This is the desired representation, the existence of which is ensured a priori by Theorem 15. From it also follows the structure of the spectrum.

We have $E_W(u) = D_{2\sqrt{u}}$. To this resolution of the identity corresponds the transformation

$$A_W = \int_0^\infty u \, d_u D_{2\sqrt{u}}, \tag{4.2.5}$$

whose domain of definition is that set of functions $x(t) \in L_2$ for which

$$\int_0^\infty u^2 \, d_u(D_{2\sqrt{u}} x, x) < \infty.$$

Using (4.2.3) once more, we see that the integral equals

$$\frac{1}{16} \int_{-\infty}^\infty t^4 \, | X(t) |^2 \, dt = \frac{1}{16} \int_{-\infty}^\infty | x''(t) |^2 \, dt.$$

Hence the domain of existence of A_W is the sub-set of L_2 which consists of the functions which are twice differentiable and whose first and second order derivatives also belong to L_2. Calculating the value of $(A_W x, y)$ we find that

$$A_W = -\frac{1}{4} \frac{d^2}{dt^2}. \tag{4.2.6}$$

Since $A_W = -\log W$, we have formally

$$W = \exp\left[\frac{1}{4} \frac{d^2}{dt^2}\right], \tag{4.2.7}$$

which is, of course, valid only for a severely restricted class of functions $x(t)$. But the corresponding inversion formula for the Gauss-Weierstrass integral, viz.

$$x(t) = \exp\left[-\frac{1}{4}\frac{d^2}{dt^2}\right] Wx(t),$$

figures in the literature and is not quite devoid of interest.[20]

In the case of Poisson's integral for the half-plane

$$P_\zeta x(t) = \frac{\zeta}{\pi}\int_{-\infty}^{\infty} \frac{x(t+u)}{\zeta^2+u^2}\,du \tag{4.2.8}$$

we arrive at similar formulas. We find that

$$(P_\zeta x, x) = \int_{-\infty}^{\infty} e^{-\zeta|s|}\,|X(s)|^2\,ds.$$

This formula shows that P_ζ is a positive definite, self-adjoint transformation whose spectrum fills out the interval [0, 1]. It also gives the canonical representation

$$(P_\zeta x, x) = \int_0^\infty e^{-\zeta u}\,d_u(D_u x, x). \tag{4.2.9}$$

Thus $E_P(u) = D_u$ and corresponds to the transformation

$$A_P = \int_0^\infty u\,d_u D_u. \tag{4.2.10}$$

A simple calculation shows that the domain of existence of A_P is that sub-set of L_2 which consists of the absolutely continuous functions whose first derivatives also belong to L_2. Using the fact that the inverse Fourier transform of $|x|\,X(t)$ is the derivative of the conjugate function of $x(t)$ which is equivalent to the conjugate of the derivative, we find that

$$A_P x = \tilde{x}'. \tag{4.2.11}$$

We note that $A_P^2 = 4A_W$ almost everywhere in the common domain of definition of the two transformations A_P^2 and A_W.

So far we have taken $E = L_2$, but W_ζ and P_ζ define semi-groups for the right half-plane in any Lebesgue space $L_p(-\infty, \infty)$, $1 \leqq p \leqq \infty$. The Dirichlet transformation is a bounded transformation on L_p to L_p for $1 < p < \infty$, but neither for $p = 1$ nor for $p = \infty$. As a consequence of this fact we find that the representations

$$W_\zeta = \int_0^\infty e^{-\zeta u}\,dD_{2\sqrt{u}}, \qquad P_\zeta = \int_0^\infty e^{-\zeta u}\,dD_u \tag{4.2.12}$$

[20] See some remarks in E. Hille [1], p. 428, concerning the validity of this formula and references to the papers of Eddington and C. W. Oseen.

remain valid for $1 < p < \infty$, but for $p = 1$ or ∞ we have to integrate by parts which brings in kernels of the Fejér type.

4.3. Semi-Groups and Abel Summability

We encounter semi-groups in the various applications of Abel's method of summability and its generalizations. We shall consider some instances.

We start with applications to sequences. The elements of the space E are then vectors $X = (x_0, x_1, \cdots, x_n, \cdots)$. We suppose that the space and its metric are such that if $\{a_n\}$ is any sequence of complex numbers such that $\sup |a_n| < \infty$, then $X \in E$ implies that $(a_n x_n) \in E$ and

$$\| (a_n x_n) \| \leqq \sup |a_n| \, \| (x_n) \| . \tag{4.3.1}$$

Let $\{\lambda_n\}$ be an arbitrary sequence of non-negative real numbers and define

$$T_\zeta X = (e^{-\zeta\lambda_n} x_n). \tag{4.3.2}$$

Then $\{T_\zeta\}$ is a semi-group in the right half-plane. In the classical case of Abel, $\lambda_n = n$, and in the various generalizations of Abel summability it is customary to assume that $\lambda_n < \lambda_{n+1}$ and $\lambda_n \to \infty$.

It is easily seen that the spectrum $S(\alpha)$ of T_α consists of the points $e^{-\alpha\lambda_n}$, $n = 0, 1, 2, \cdots$, and their limit points. Thus the spectrum is real positive and restricted to the interval $[0, 1]$. It is easy to verify that T_α is uniformly continuous for $\alpha > 0$, whence it follows that (T_3) and (T_4) are satisfied. A simple calculation shows that (T_6') also holds with $K(\alpha) = 1$ and $\pi(\alpha) = 1$. Thus all the assumptions of Theorem 13 hold and formula (3.5.14) holds with $\beta > 1$, $M = 1$. Actually we can take $\beta \geqq 0$. Indeed we have

$$T_\zeta = \int_0^\infty e^{-\zeta u} \, dA_0(u),$$

where $A_0(u)$ is the evidently bounded operation which consists in replacing the vector $(x_0, x_1, \cdots, x_n, \cdots)$ by the vector $(x_0, x_1, \cdots, x_{[u]}, 0, 0, \cdots)$. Here $[u]$ denotes the integral part of u, except when $u = 0$, when all components are to be replaced by 0.

We get somewhat more interesting formulas in the case of function spaces. Let $E = L_p(-\pi, \pi)$, $1 \leqq p < \infty$, with the customary metric. Let $x(t) \in L_p$ and put

$$x(t) \sim \sum_{-\infty}^{\infty} x_n e^{nit}. \tag{4.3.3}$$

Let us form

$$T_\zeta x(t) \sim \sum_{-\infty}^{\infty} x_n e^{-\zeta\lambda_{|n|}} e^{nit}, \tag{4.3.4}$$

where $\Re(\zeta) > 0$. The notation suggests that the series on the right is a Fourier series, but this is by no means the case unless the λ_n-sequence is subject to

special conditions. Necessary and sufficient conditions in order that T_ζ shall define a bounded transformation on L_p to L_p for every ζ in the right half-plane are given by the theory of factor sequences of Fourier series (see A. Zygmund [1], pp. 100–105). In order to simplify matters we suppose that λ_n is monotone increasing and

$$\lambda_n/\log n \to +\infty \quad \text{as} \quad n \to \infty.$$

These conditions are sufficient in order to ensure that T_ζ shall have the desired property.

Under these circumstances T_ζ defines a semi-group in the right half-plane. The canonical form of T_ζ is easily obtained from (4.3.4). We define an operator $A_0(u)$ which carries the Fourier series (4.3.3) into the N^{th} partial sum of the series where N is the largest value of n for which $\lambda_n \leqq u$. Then

$$T_\zeta = \int_0^\infty e^{-\zeta u}\, dA_0(u).$$

Take the simplest case of all, $p = 2$, $\lambda_n = n$, i.e.,

$$\Pi_\zeta x(t) = \sum_{-\infty}^{\infty} x_n e^{-|n|\zeta} e^{nit}, \tag{4.3.5}$$

which leads to Poisson's integral for the circle if we put $e^{-\zeta} = r$. $A_0(u)$ is then simply $E_\Pi(u)$, the resolution of the identity of Theorem 15, and a simple calculation shows that the corresponding operator A_Π is defined by

$$(A_\Pi x, y) = \sum_{n=-\infty}^{\infty} |n|\, x_n \bar{y}_n.$$

It follows that

$$A_\Pi x = \tilde{x}', \tag{4.3.6}$$

i.e., we arrive at formally the same operation as in the case of Poisson's integral for the half-plane, see formula (4.2.11). Actually the operators, though closely related, are completely distinct.

We can also describe A_Π as the operator which multiplies the n^{th} Fourier coefficient of $x(t)$ by $|n|$. This property is characteristic for the Poisson-Abel transformation and is independent of the coördinate system, i.e., of the orthogonal system. Let (a, b) be an arbitrary finite or infinite interval, $\{\omega_n(t)\}$ an ortho-normal system, complete in $L_2(a, b)$. Let $x(t) \in L_2(a, b)$ and

$$x(t) \sim \sum_{n=1}^{\infty} x_n \omega_n(t). \tag{4.3.7}$$

Define

$$Q_\zeta x(t) \sim \sum_{n=1}^{\infty} x_n e^{-n\zeta} \omega_n(t). \tag{4.3.8}$$

This is again a semi-group defined in the right half-plane and admits of a representation of the form

$$Q_\zeta = \int_0^\infty e^{-\zeta u}\, dE_Q(u)$$

Here $E_Q(u)$ is the operator which replaces the orthogonal series (4.3.7) by its N^{th} partial sum where $N = [u]$. The corresponding operator A_Q carries the series (4.3.7) into the series

$$\sum_{n=1}^{\infty} n x_n \omega_n(t),$$

i.e., multiplies the n^{th} Fourier coefficient of $x(t)$ by n.

The case of Hermitian series is particularly interesting. We have $a = -\infty$, $b = +\infty$, and $\omega_n(t)$ equals

$$\eta_n(t) = \{\pi^{\frac{1}{2}} 2^n n!\}^{-\frac{1}{2}} H_n(t) e^{-t^2/2}. \tag{4.3.9}$$

Denote the corresponding Abel-Poisson operator by H_ζ, and let A_H denote $-\log H_1$ as usual. It turns out that A_H is the differential operator

$$A_H x(t) = -\tfrac{1}{2}[x''(t) + (1 - t^2)x(t)] \tag{4.3.10}$$

which enters in so many other problems involving Hermitian orthogonal functions. It is enough to recall that

$$A_H \eta_n(t) = n\eta_n(t). \tag{4.3.11}$$

This relation also reveals how (4.3.10) can be proved.

These examples could be multiplied ad infinitum, but the instances already given are enough to show the possibilities of obtaining formulas of varying degree of interest.

4.4. Fractional Integration

The best known functional operation having the semi-group property is probably the Riemann-Liouville fractional integral.

Let E be a Lebesgue space $L_p(0, a)$, $1 \leqq p \leqq \infty$, $a < \infty$, and form

$$I^\zeta f = \frac{1}{\Gamma(\zeta)} \int_0^t (t - u)^{\zeta - 1} f(u)\, du \equiv f_\zeta(t), \tag{4.4.1}$$

where $\Re(\zeta) > 0$. It is well known that $f_\zeta(t) \in L_p(0, a)$ and that $I^\zeta f$ has the semi-group property in the right half-plane. Stirling's formula for the gamma function shows that I^ζ, which is a holomorphic operator function in the right half-plane, is bounded in the sense of formula (2.5.3) in the half-plane. Indeed, we find $\gamma(\xi) = \pi/2$ [see formula (2.3.12)] whence it follows that $\omega_0 = \omega_1 = 1$. Consequently there exist representations by means of binomial series in the variable ζ/ω for $0 < \omega < 1$ which are weakly convergent for $\Re(\zeta) > 0$ and strongly as well as uniformly convergent at least for $\Re(\zeta) > \omega$.

But in the present case we have convergent representations by binomial

series in terms of the variable ζ itself. We get the most elegant representation in terms of the variable $\zeta - 1$, however. For this purpose we use the known series[21]

$$\frac{v^{\zeta-1}}{\Gamma(\zeta)} = \sum_{n=0}^{\infty} (-1)^n L_n(v) \binom{\zeta - 1}{n}.$$

Here $L_n(v)$ denotes the n^{th} polynomial of Laguerre. When $v > 0$ the series converges for $\xi > \frac{1}{4}$ and absolutely for $\xi > \frac{3}{4}$. When $v = 0$ we must have $\xi > 1$ and the series is absolutely convergent. With the aid of this series we obtain the expansion

$$I^{\zeta} f = \sum_{n=0}^{\infty} (-1)^n \binom{\zeta - 1}{n} \int_0^t L^n(t - u) f(u)\, du. \tag{4.4.2}$$

Using the well known estimate

$$| L_n(t) | \leqq e^{t/2},$$

we see that the series converges strongly as well as uniformly at least for $\Re(\zeta) > 1$.

The resolvent of I^{α} is known. We have[22]

$$R(\lambda, \alpha) f = \frac{1}{\lambda} \frac{d}{dt} \int_0^t E_{\alpha}[(t - u)^{\alpha}/\lambda] f(u)\, du, \tag{4.4.3}$$

where

$$E_{\alpha}(z) = \sum_{n=0}^{\infty} \frac{z^n}{\Gamma(n\alpha + 1)}$$

is the Mittag-Leffler function.

This formula shows that $R(\lambda, \alpha)$ is an entire function of $1/\lambda$ for every $\alpha > 0$, so that the spectrum reduces to just one point, namely $\lambda = 0$. Supposing $0 < \alpha < 2$, let V_{α} denote the sector $\frac{\alpha}{2}\pi < \arg z < \left(2 - \frac{\alpha}{2}\right)\pi$ and use the same notation in the λ-plane. From the classical estimate

$$E_{\alpha}(z) = \frac{1}{\Gamma(-\alpha) z} + O\left(\frac{1}{|z|^2}\right),$$

valid for z in V_{α}, we conclude that

$$\| R(\lambda, \alpha) \| \leqq M(\alpha) | \lambda |^{-1}, \qquad \lambda \in V_{\alpha}. \tag{4.4.4}$$

In the complementary sector we have

$$\| R(\lambda, \alpha) \| \leqq M(\alpha) | \lambda |^{-1-1/\alpha} \exp [a | \lambda |^{-1/\alpha}]$$

and this estimate is not capable of essential improvement.

---

[21] Due to S. Wigert [1], p. 18. Cf. N. E. Nörlund [2], pp. 161–165.

[22] See E. Hille and J. D. Tamarkin [1], pp. 524–525.

It follows from these estimates that the spectrum of I^α satisfies the assumptions of Theorem 9 for $0 < \alpha < 1$.

Finally let us consider an example which does not fit into the general theory of semi-groups developed in this paper. In a recent important memoir M. Riesz [1] has considered a functional transformation of the Riemann-Liouville type. Let Ω_m be a Euclidean space of $m \geqq 1$ dimensions. Let r_{PQ} denote the distance between the two points P and Q and form

$$J^\alpha f(P) = \frac{\Gamma\left(\frac{m-\alpha}{2}\right)}{\pi^{m/2}2^\alpha\Gamma\left(\frac{\alpha}{2}\right)} \int_{\Omega_m} f(Q) r_{PQ}^{\alpha-m}\, dQ, \tag{4.4.5}$$

where the only assumption on f is that the integral shall be absolutely convergent for all finite P in Ω_m. Riesz takes α to be positive and shows that

$$J^\alpha[J^\beta f(P)] = J^{\alpha+\beta} f(P), \qquad \alpha > 0, \qquad \beta > 0, \qquad \alpha + \beta < m. \tag{4.4.6}$$

Formula (4.4.5) also has a sense for complex values of α. Indeed, if $J^\xi f(P)$ has a sense for $\zeta = \xi$ then it also exists for $\zeta = \xi + i\eta$, where η is arbitrary. Formula (4.4.6) also holds if we suppose the real parts of the exponents restricted to the interval $(0, m)$. Thus we are dealing with an operator family having the semi-group property in a strip of the complex plane. But it is clear that the operator J^α as a rule does not define a bounded transformation on and to a Banach space, so that our theory does not apply.

Appendix

5.1. Carlson's Theorem

We shall state and prove that part of Theorem E, due to F. Carlson, which was used in §3.3.

Theorem E. *Let $f(z)$ be holomorphic in $x \geqq 0$. In addition suppose that* (i) *on the imaginary axis*

$$|f(\pm ir)| \leqq Ce^{\pi r}$$

(ii) *for* $-\pi/2 \leqq \theta \leqq \pi/2$

$$|f(re^{i\theta})| \leqq Ce^{\lambda(\theta)r}, \qquad \lambda(\theta) \leqq M,$$

where $\lambda(-\theta) = \lambda(\theta)$ *and*

$$\limsup_{\delta\to 0} \frac{1}{\delta}\left[\pi - \lambda\left(\frac{\pi}{2} - \delta\right)\right] = +\infty,$$

and (iii) $f(n) = 0$ *for* $n = 1, 2, 3, \cdots$. *Then* $f(z) \equiv 0$.

Proof. Let $\omega > 0$ and form

$$H(z, \omega) = zf(z)e^{\omega z}/\sin \pi z.$$

This function is holomorphic for $x \geqq 0$. Take a ray $\arg z = \pi/2 - \delta$. Then

$$|H(z,\omega)| \leqq B \exp\left\{r\left[\lambda\left(\frac{\pi}{2}-\delta\right)+\omega\sin\delta-\pi\cos\delta\right]\right\}$$

$$= B\exp\left\{-r\sin\delta\left[\frac{\pi-\lambda\left(\frac{\pi}{2}-\delta\right)}{\sin\delta}-\omega-\pi\frac{1-\cos\delta}{\sin\delta}\right]\right\}$$

Assumption (ii) shows that for fixed ω and suitably chosen small positive δ the expression inside the second square brackets can be made as large positive as we please. Hence we can choose a sequence $\delta_n \to 0$ and find numbers $\omega_n \to \infty$ as $n \to \infty$ such that

$$|H(re^{\pm i\theta_n},\omega)| \leqq B$$

for $\theta_n = \pi/2 - \delta_n$, $0 \leqq \omega \leqq \omega_n$.

In the sector $V_n : |\arg z| \leqq \theta_n$, we can find a constant A_n such that $|H(z,\omega)| \leqq M\exp(A_n r)$, $z \in V_n$, $0 \leqq \omega \leqq \omega_n$. An application of the classical Phragmén-Lindelöf theorem to the function $H(z,\omega)$ in the sector V_n shows that

$$|H(z,\omega)| \leqq B, \qquad z \in V_n, \qquad 0 \leqq \omega \leqq \omega_n.$$

But this implies that $|H(z,\omega)| \leqq B$ for $x \geqq 0$ and all values of ω. Hence

$$|zf(z)| \leqq B|\sin\pi z|e^{-\omega x}$$

for all $\omega > 0$. But this means that $|f(z)|$ is arbitrarily small for $x \geqq \epsilon > 0$. Consequently $f(z) \equiv 0$.

5.2 Extensions of Theorem 4[23]

In the present paragraph we shall extend Theorem 4 of §2.4 in three different directions. We shall show that a semi-group whose norm is measurable need not have a finite least upper bound on the interval $(0, a)$. On the other hand, it does have a positive greatest lower bound which tends to a limit $\geqq 1$ as $a \to 0$. Finally, a group whose norm is measurable has a finite least upper bound on any finite interval $(-a, a)$.

THEOREM 4_1. *There exists a space E of type (B) and an operator function $W(\zeta)$ on E to E which is holomorphic and has the semi-group property for $\Re(\zeta) > 0$, such that $\|W(\zeta)\| \to \infty$ as $\zeta \to 0$ along the positive real axis.*

PROOF. We can take for E the space $L(-\pi, \pi)$. We shall use a modification of the transformations defined by formula (4.3.4) obtained by introducing gaps in the factor sequence. Let $\{p_k\}$ be a monotone increasing sequence of positive integers such that $\sum 1/p_k$ is divergent, and form the function

$$F(t;\zeta) = \sum_{k=1}^{\infty} e^{-p_k\zeta}\cos p_k t, \qquad \Re(\zeta) > 0. \tag{5.2.1}$$

[23] Added in proof October 22, 1938.

If $x(t) \in L(-\pi, \pi)$ we define

$$W(\zeta)[x(t)] = (2\pi)^{-1} \int_{-\pi}^{\pi} F(t-u;\zeta)x(u)\,du. \tag{5.2.2}$$

It is a simple matter to show that this is a bounded transformation on L to L whose norm equals

$$\| W(\zeta) \| = \pi^{-1} \int_0^{\pi} | F(t;\zeta) |\,dt. \tag{5.2.3}$$

The transformation is evidently holomorphic and has the semi-group property in the right half plane.

We shall now impose the additional restriction that $p_k \equiv 1 \pmod 4$ for every k. Then

$$\begin{aligned}
\pi \| W(\xi) \| &> \int_0^{\pi/2} | F(t;\xi) |\,dt \\
&= \int_0^{\pi/2} \Big| \sum_1^{\infty} e^{-p_k\xi} - \sum_1^{\infty} [1 - \cos p_k t] e^{-p_k\xi} \Big| dt \\
&\geqq (\pi/2) \sum_1^{\infty} e^{-p_k\xi} - \sum_1^{\infty} [(\pi/2) - (1/p_k)] e^{-p_k\xi} \\
&= \sum_1^{\infty} e^{-p_k\xi}/p_k > e^{-1} \sum_1^{[1/\xi]} 1/p_k,
\end{aligned}$$

and this expression tends to $+\infty$ as $\xi \to 0$.

The existence of such functions shows that the boundedness of $\| T_\alpha \|$ on an interval $(0, a)$ is not a consequence of the semi-group property and the measurability of $\| T_\alpha \|$. Thus Theorem 4 is the best of its kind.

THEOREM 4_2. *If T_α has the semi-group property and $\| T_\alpha \|$ is measurable, both for $\alpha > 0$, and if $\lambda(a) = \inf \| T_\alpha \|$ when $0 < \alpha \leqq a$, then $\lambda(a)$ is positive for every a and increases to a limit $\geqq 1$ as a decreases to zero, unless T_α is identically zero for $\alpha > 0$.*

PROOF. It is clear that $\lambda(a) \geqq 0$ is a never-increasing function of a. Suppose that $\lambda(a_0) = 0$. There is then a sequence $\{\alpha_n\}$ tending to a_0 such that, in the notation of §2.4, $e(\alpha_n) \leqq -n$. By Theorem 4 we can find a finite M such that $e(\beta) < M$ for $1 \leqq \beta \leqq 2$. Formula (2.4.7) shows that for such β's we have $e(\alpha_n + \beta) \leqq -n + M$. We conclude from this the existence of a fixed measurable set of measure $\geqq 1$ in which the inequality $e(\alpha) < -n + M$ holds for every n. This implies $e(\alpha) = -\infty$ in the set in question. This is impossible unless T_α is identically zero for every $\alpha > 0$. We have consequently $\lambda(a) > 0$ for every a. Using (2.4.7) once more, and putting $\liminf_{\alpha \to 0} e(\alpha) = e_0$, we see that $e_0 \leqq 2e_0$ or $e_0 \geqq 0$. This completes the proof of the theorem.

THEOREM 4_3. *If T_α defines a group on $(-\infty, \infty)$ and $\| T_\alpha \|$ is measurable, then $\| T_\alpha \|$ is bounded on every finite interval $(-a, a)$.*

PROOF. We proceed as in the proof of Theorem 4. If the theorem is false we can find a sequence $\alpha_n \to \gamma$ such that $p(\alpha_n) \geqq 2n$. We let α and β be arbi-

trary numbers from the interval $(-a, a)$ such that $\alpha + \beta = \alpha_n$. We can then conclude that $p(\alpha) > n$ for every n on a fixed measurable α-set of positive measure. Hence $p(\alpha) = +\infty$ on this set. This being impossible, $p(\alpha)$ must be bounded on $(-a, a)$ and consequently also $\| T_\alpha \|$.

REFERENCES

BANACH, S. [1] Théorie des opérations linéaires. Monografje mat. I. Warsaw, 1932.

BERNSTEIN, S. [1] Sur les fonctions absolument monotones. Acta Mathematica, 52 (1929) 1–66.

CARLSON, F. [1] Sur une classe de série de Taylor. Thèse, Upsal, 1914, 76 pp.

[2] Sur les séries de coefficients binomiaux. Nova Acta R. Soc. Sci. Upsaliensis, (4) 4 (1915), No. 3, 61 pp.

DUNFORD, N. [1] On one-parameter groups of linear transformations. Annals of Mathematics, (2) 39 (1938) 569–573.

HAMEL, G. [1] Eine Basis aller Zahlen und unstetiger Lösungen der Funktionalgleichung $f(x + y) = f(x) + f(y)$. Math. Annalen, 60 (1905) 459–462.

HAUSDORFF, F. [1] Momentprobleme für ein endliches Intervall. Math. Zeitschrift, 16 (1923) 220–248.

HILLE, E. [1] A class of reciprocal functions. Annals of Mathematics, (2) 27 (1926) 427–464.

[2] Notes on linear transformations. I. Trans. Amer. Math. Soc., 39 (1936), 131–153.

[3] On semi-groups of transformations in Hilbert space. Proc. Nat. Acad. Sci., 24 (1938) 159–161.

HILLE, E. & TAMARKIN, J. D. [1] On the theory of linear integral equations. I. Annals of Mathematics, (2) 31 (1930) 479–528.

[2] On the theory of Laplace integrals. Proc. Nat. Acad. Sci., 19 (1933) 908–912; 20 (1934) 140–144.

LINDELÖF, E. [1] Le calcul des résidus et ses applications à la théorie des fonctions. Paris, Gauthier-Villars, 1905.

NEUMANN, J. VON [1] Zur Algebra der Funktionaloperationen und Theorie der normalen Operatoren. Math. Annalen, 102 (1929) 370–427.

NÖRLUND, N. E. [1] Sur les formules d'interpolation de Stirling et de Newton. Ann. Ecole Norm. Sup., (3) 39 (1922) 343–403; 40 (1923) 35–54.

[2] Leçons sur les séries d'interpolation. Paris, Gauthier-Villars, 1926.

PINCHERLE, S. [1] Sur les fonctions déterminantes. Ann. Ecole Norm. Sup., (3) 22 (1905) 9–68.

RIESZ, F. [1] Über die linearen Transformationen des komplexen Hilbertschen Raumes. Acta Litt. ac Sci., Szeged, 5 (1930) 23–54.

RIESZ, M. [1] Intégrales de Riemann-Liouville et potentiels. Acta. Litt. ac Sci., Szeged, 9 (1938) 1–42.

STONE, M. H. [1] Linear transformations in Hilbert space and their applications to analysis. Amer. Math. Soc. Coll. Publ. XV. New York, 1932.

SZ. NAGY, B. VON [1] Über messbare Darstellungen Liescher Gruppen. Math. Annalen, 112 (1936) 268–296.

TAYLOR, A. E. [1] Linear operations which depend analyticaliy upon a parameter. Annals of Mathematics, (2) 39 (1938) 574–593.

WIDDER, D. V. [1] Necessary and sufficient conditions for the representation of a function as a Laplace integral. Trans. Amer. Math. Soc., 33 (1931) 851–892.

WIENER, N. [1] Note on a paper of M. Banach. Fundamenta Mathematicae, 4 (1923) 136–143.

WIGERT, S. [1] Contributions à la théorie des polynomes d'Abel-Laguerre. Arkiv f. Mat., 15 (1921) no. 25, 22 pp.

ZYGMUND, A. [1] Trigonometrical series. Monografje Mat., V. Warsaw, 1935.

ANNALS OF MATHEMATICS
Vol. 40, No. 1, January, 1939

ON SOME PROPERTIES OF RECTILINEAR CONGRUENCES

BY HAROLD HILTON

(Received October 19, 1937, Revised June 16, 1938)

1. Consider a "director surface" any one of whose points is given by parameters u, v and whose fundamental differential forms are

$$E\,du^2 + 2F\,du\,dv + G\,dv^2, \qquad L\,du^2 + 2M\,du\,dv + N\,dv^2.$$

Suppose that through any one of its points (x, y, z) a line passes with direction cosines λ, μ, ν making angles α, β with the tangents to $v = \text{const.}$, $u = \text{const.}$ If we denote partial differentiation with respect to u and v by suffixes 1 and 2 and

$$a \equiv \lambda x_1 + \mu y_1 + \nu z_1, \qquad b \equiv \lambda x_2 + \mu y_2 + \nu z_2, \tag{1}$$

then

$$\sqrt{E}\cos\alpha = a, \qquad \sqrt{G}\cos\beta = b. \tag{2}$$

Hence a knowledge of a, b and the director surface completely determines the rectilinear congruence formed by the line. In §§1, 2, 3 we shall exclude the case in which the director surface touches the line and is a sheet of the focal surface of the congruence. We shall therefore have $a^2 < E$ and $b^2 < G$.

We shall suppose throughout that $\sqrt{(EG - F^2)}$ is taken positive, shall denote the determinant

$$\begin{vmatrix} \lambda & \mu & \nu \\ x_1 & y_1 & z_1 \\ x_2 & y_2 & z_2 \end{vmatrix}$$

by $|\,\lambda y_1 z_2\,|$ etc., etc., and shall adopt the abbreviations

$$E_0 \equiv E - a^2, \qquad F_0 \equiv F - ab, \qquad G_0 \equiv G - b^2. \tag{3}$$

If ϕ is the angle between the ray of the congruence and the normal to the director surface,

$$\sqrt{(EG - F^2)}\cos\phi = |\,\lambda y_1 z_2\,|, \tag{4}$$

so that

$$(EG - F^2)\cos^2\phi = (E_0 G_0 - F_0^2) \tag{5}$$

by the ordinary rule for squaring a determinant.

From (i) and (iv) we get

$$\lambda(EG - F^2) = \begin{vmatrix} X\cos\phi & -a & -b \\ x_1 & E & F \\ x_2 & F & G \end{vmatrix}, \text{ etc.} \tag{6}$$

where X, Y, Z are the direction cosines of the normal to the director surface, so that

$$X\sqrt{(EG - F^2)} = y_1 z_2 - y_2 z_1, \text{ etc.}$$

The focal and limiting points, etc. of a congruence are usually given by formulae involving $\mathfrak{E}$, $\mathfrak{F}$, $\mathfrak{G}$, e, f, f', g defined by

$$\mathfrak{E} \equiv \lambda_1^2 + \mu_1^2 + \nu_1^2, \quad \mathfrak{F} \equiv \lambda_1\lambda_2 + \cdots, \quad \mathfrak{G} \equiv \lambda_2^2 + \cdots,$$

$$e \equiv \lambda_1 x_1 + \cdots, \quad f \equiv \lambda_1 x_2 + \cdots, \quad f' \equiv \lambda_2 x_1 + \cdots, \quad g = \lambda_2 x_2 + \cdots.$$

These quantities are connected with a, b by the formulae

$$\begin{cases} e &= a_1 - \{11, 1\}a - \{11, 2\}b - L\cos\phi, \\ f &= b_1 - \{12, 1\}a - \{12, 2\}b - M\cos\phi, \\ f' &= a_2 - \{12, 1\}a - \{12, 2\}b - M\cos\phi, \\ g &= b_2 - \{22, 1\}a - \{22, 2\}b - N\cos\phi, \end{cases} \tag{7}$$

$$\begin{cases} (E_0G_0 - F_0^2)\mathfrak{E} = E_0f^2 - 2F_0ef + G_0e^2, \\ (E_0G_0 - F_0^2)\mathfrak{F} = E_0fg - F_0(eg + ff') + G_0ef', \\ (E_0G_0 - F_0^2)\mathfrak{G} = E_0g^2 - 2F_0gf' + G_0f'^2, \end{cases} \tag{8}$$

where $\cos\phi$ is given by (4) and (5) and

$$\{11, 1\}, \quad \{12, 1\}, \quad \{22, 1\}, \quad \{11, 2\}, \quad \{12, 2\}, \quad \{22, 2\}$$

are the familiar Christoffel symbols whose values are obtained on dividing by $2(EG - F^2)$ the quantities

$$GE_1 + FE_2 - 2FF_1, \quad GE_2 - FG_1, \quad -FG_2 - GG_1 + 2GF_2,$$

$$-FE_1 - EE_2 + 2EF_1, \quad EG_1 - FE_2, \quad EG_2 + FG_1 - 2FF_2$$

respectively.

The first of the formulae (7) is got on multiplying $|\lambda y_1 z_2|$, $|x_1 y_1 z_2|$ by the ordinary rule for multiplying determinants. The first of (8) is got on forming the squares and product of $|\lambda y_1 z_2|$, $|\lambda_1 y_1 z_2|$; and similarly for the other formulae.

We note in passing that we can obtain conversely a, b, E, F, G, L, M, N in terms of $\mathfrak{E}$, $\mathfrak{F}$, $\mathfrak{G}$, e, f, f', g and their derivatives with respect to u and v, so that the director surface and congruence are obtainable when we know $\mathfrak{E}$, $\mathfrak{F}$, $\mathfrak{G}$, e, f, f', g.

For two equations giving a and b are got on forming the products of each of $|\lambda\mu_1\nu_2|$, $|\lambda_{11}\mu_1\nu_2|$, $|\lambda_{12}\mu_1\nu_2|$, $|\lambda_{22}\mu_1\nu_2|$ with each of $|\lambda\mu_1\nu_2|$, $|x_1\mu_1\nu_2|$, $|x_2\mu_1\nu_2|$ and combining the results suitably.

Then E, F, G are got on forming the squares of each of $|\lambda\mu_1\nu_2|$, $|x_1\mu_1\nu_2|$, $|x_2\mu_1\nu_2|$ and the product of any two. To get L, M, N multiply each of

$$|\lambda y_1 z_2|, \quad |x_{11} y_1 z_2|, \quad |x_{12} y_1 z_2|, \quad |x_{22} y_1 z_2| \quad \text{by} \quad |\lambda y_1 z_2|.$$

We show readily that

$$eg - ff' = \sqrt{(\mathfrak{E}\mathfrak{G} - \mathfrak{F}^2)}\sqrt{(EG - F^2)}\cos\phi.$$

Since we have supposed that the ray does not touch the surface, $eg \neq ff'$. With this restriction we obtain the equation

$$\begin{vmatrix} e\,du + f'\,dv & f\,du + g\,dv \\ E_0\,du + F_0\,dv & F_0\,du + G_0\,dv \end{vmatrix} = 0 \tag{9}$$

for the developables of the congruence on substituting for $\mathfrak{E}$, $\mathfrak{F}$, $\mathfrak{G}$ from (8) in the well-known equation

$$\begin{vmatrix} e\,du + f\,dv & f'\,du + g\,dv \\ \mathfrak{E}\,du + \mathfrak{F}\,dv & \mathfrak{F}\,du + \mathfrak{G}\,dv \end{vmatrix} = 0.$$

Similarly we obtain

$$\begin{vmatrix} er + E_0 & fr + F_0 \\ f'r + F_0 & gr + G_0 \end{vmatrix} = 0 \tag{10}$$

from the familiar equation

$$\begin{vmatrix} \mathfrak{E}r + e & \mathfrak{F}r + f \\ \mathfrak{F}r + f' & \mathfrak{G}r + g \end{vmatrix} = 0$$

for the distance r of the foci of the ray from the director surface (measured along the congruence's ray).

The condition $f = f'$ for a normal congruence gives $a_2 = b_1$ (see L. Bianchi, *Geometria Differenziale*, vol. 1 (Pisa, 1922), p. 475) and the conditions

$$e : \tfrac{1}{2}(f + f') : g = \mathfrak{E} : \mathfrak{F} : \mathfrak{G}$$

for an isotropic congruence give

$$e : \tfrac{1}{2}(f + f') : g = E_0 : F_0 : G_0\,. \tag{11}$$

2. We shall now take a few illustrations of the use of the formulae of §1.

Take on a plane $x = u$, $y = v$, $z = 0$. Then

$$E = G = 1, \quad F = L = M = N = 0; \tag{1}$$
$$e = a_1\,, \quad f = b_1\,, \quad f' = a_2\,, \quad g = b_2\,.$$

Suppose that the plane is the mean surface of an isotropic congruence, so that $2e = f + f' = 2g = 0$. Then either a and b are both constant so that the rays of the congruence are all parallel, or else with a suitable choice of origin $a = v/k$, $b = -u/k$ (k constant). In this latter case the rays of the congruence are the generators of one-sheeted hyperboloids of revolution with a common centre and axis, the radii of whose waists are proportional to the sines of the semi-vertical angle of their asymptotic cones.

If the plane is the mean surface of a normal congruence, we have $f = f'$ and $E_0 g - F_0(f + f') + G_0 e = 0$ by §1 (10). Therefore

$$a_2 = b_1, \qquad (1 - a^2)b_2 + ab(a_2 + b_1) + (1 - b^2)a_1 = 0.$$

These equations give us the following method of getting a and b. Take any minimal surface whatever $z = \psi(x, y)$; then

$$a = \frac{\partial}{\partial x}\psi(xi, yi), \qquad b = \frac{\partial}{\partial y}\psi(xi, yi).$$

For instance, from Scherk's minimal surface

$$z = c \log \{\sec(x/c) \cos(y/c)\}$$

we get

$$a = -\tanh(x/c), \qquad b = \tanh(y/c).$$

A general discussion of congruences with a plane mean surface will be found in a paper by P. Vincensini, Annales Fac. Sciences Toulouse, III 19 (1927), p. 93.

Suppose again that the two families of developables of a congruence meet a plane in two series of mutually perpendicular parallel lines.

Taking these lines as $x = \text{const.}$, $y = \text{const.}$ we have from §1 (9) $eF_0 = fE_0$, $gF_0 = f'G_0$ leading to

$$a^2 = (V - 1)/(UV - 1), \qquad b^2 = (U - 1)/(UV - 1),$$

where U is any function of u only and V of v.

As another illustration take the case of a right circular cone with vertical angle 2γ

$$x = u \cos v, \qquad y = u \sin v, \qquad z = u \cot \gamma.$$

Then

$$E = \operatorname{cosec}^2 \gamma, \qquad G = u^2, \qquad N = u \sec \gamma, \qquad F = L = M = 0.$$

Straightforward work shows that the equations $2e = f + f' = 2g = 0$ lead to

$$a = -\cot \gamma \sin \delta + \cos \delta \sin v, \qquad b = u \cos \delta \cos v,$$

where δ is some constant.

From §1 (6), or otherwise, we see that $\lambda = 0$, $\mu = \cos \delta$, $\nu = -\sin \delta$. The case of the right circular cylinder is handled similarly, and we have: The only

isotropic congruence having a right circular cone or a right circular cylinder as its mean surface is the congruence consisting of parallel lines.

Lastly consider the developable which is formed by tangents to the locus of (ξ, η, ζ), u being its arc, ρ and σ its radii of curvature and torsion, and v the distance measured along the tangent from (ξ, η, ζ). Then

$$(2) \quad E = 1 + (v/\rho)^2, \qquad F = G = 1, \qquad L = v/(\rho\sigma), \qquad M = N = 0.$$

Let us find the congruence which has the developable as a sheet of the limiting surface (locus of limiting points), each generator cutting at the same angle all the rays of the congruence which it meets. We have $g = 0, f + f' = 0$, which give us $b_2 = 0, v(a_2 + b_1) - 2a + 2b = 0$, so that

$$a = U + U_1 v + \mathfrak{U} v^2, \qquad b = U,$$

U and $\mathfrak{U}$ being any functions of u.

The rays meeting any given generator form one of the principal ruled surfaces of the congruence.

3. H. Horninger in Sitzungsber. Akad. Wiss. Wien, Math. Phys. Klasse, 145 (1936), p. 368 has considered the congruences obtained by reflecting at a given surface a pencil of rays proceeding from a point.

We shall now consider a somewhat similar problem: A congruence is refracted or reflected at a given surface, the ratio of the indices of refraction of the media on either side of the surface being $1:\mu$ ($\mu = -1$ for reflection). Find the congruence, if all the refracted congruences have a certain property, whatever μ may be.

If $a = b = 0$, all the refracted congruences coincide with the incident congruence. We shall suppose this case excluded.

It is readily seen from §1 (2) that the result of refracting a congruence in its director surface is to replace a and b by μa and μb. Let the given property lead to an equation of the form $\psi(\mathfrak{E}, \mathfrak{F}, \mathfrak{G}, e, f, f', g) = 0$. Then, if we substitute for $\mathfrak{E}, \mathfrak{F}, \mathfrak{G}, e, f, f', g$ in terms of a, b, E, F, G, L, M, N by means of §1 (7), (8), we shall get an equation which is to be unaltered when a and b are replaced by μa and μb, whatever μ may be. If this equation is algebraic of degree n in μ, it is necessary and sufficient that the $n + 1$ coefficient of powers of μ should all be zero in order that all refracted congruences should have the property; and, if $n + 1$ of the refracted congruences have the property, all will do so.

The simplest illustration is the well known result that, if a congruence is normal, all the refracted congruences are. For this we must have $f = f'$ and therefore $a_2 = b_1$. More generally, any property leading to an equation homogeneous in a and b is possessed by all the congruences, if it is possessed by the incident congruence (see the end of §2).

In applying the method it will be convenient to choose parameters on the surface so that $F = b = 0$, or so that $F = M = 0$ the lines of curvature being

parametric curves, or so that (ϵ being the exponential constant, since e has another meaning)

$$(1) \quad E = G = \epsilon^{2t}, \quad F = 0, \quad L = l\epsilon^{t}, \quad M = m\epsilon^{t}, \quad N = n\epsilon^{t},$$

with the Codazzi and Gauss equations

$$(2) \quad l_2 - m_1 = ml_1 + nl_2, \quad n_1 - m_2 = ll_1 + ml_2, \quad ln - m^2 + l_{11} + l_{22} = 0.$$

This last is possible in an infinite number of ways, taking any isometric net on the surface as parametric.

It must suffice here to state merely the results of straightforward applications of the method.

To see whether all refracted congruences can be isotropic, we may start with equations (1). We find that the only possibility is that discussed in §2, in which the refracting surface is plane and is the mean surface of the congruence.

To see whether all refracted congruences can have the refracting surface as mean surface we start with equations (1) and find readily that the surface must be plane or minimal.

If it is plane, we take the equations of §2 (1) and get

$$a_1 + b_2 = 0, \quad a^2b_2 + b^2a_1 = ab(a_2 + b_1).$$

These equations give $a = \psi_2$, $b = -\psi_1$,

$$(3) \quad \psi_1\psi_2(\psi_{11} - \psi_{22}) = (\psi_1^2 - \psi_2^2)\psi_{12}.$$

Thence we have

$$\psi_2 x - \psi_1 y = \psi_2\Phi(\psi_1/\psi_2),$$

which is readily solved by Legendre's substitution

$$\psi_1 = X, \quad \psi_2 = Y, \quad x\psi_1 + y\psi_2 - \psi = Z$$

followed by $X = r\cos\theta$, $Y = r\sin\theta$. We note that (3) is the differential equation of the surfaces $z = \psi(x, y)$, the projection of whose lines of curvature on $z = 0$ form an orthogonal net.

If the surface is minimal, we take it in the form

$$(4) \quad \begin{cases} x = \frac{1}{4}\int \frac{1-\theta^2}{\theta'}dw + \frac{1}{4}\int \frac{1-\Theta^2}{\Theta'}dW, \\ y = \frac{1}{4}i\int \frac{1+\theta^2}{\theta'}dw - \frac{1}{4}i\int \frac{1+\Theta^2}{\Theta'}dW, \\ z = \frac{1}{2}\int \frac{\theta}{\theta'}dw + \frac{1}{2}\int \frac{\Theta}{\Theta'}dW, \end{cases}$$

$$(5) \quad E = G = (\theta\Theta + 1)^2/(4\theta'\Theta'), \quad F = M = 0, \quad L = -1, \quad N = 1,$$

where $w \equiv u + iv$, $W \equiv u - iv$, θ is a function of w, and Θ is what θ becomes when we replace i by $-i$.

We now have $l = -\epsilon^{-t}$, $m = 0$, $n = \epsilon^{-t}$ in (1) and get

$$a_1 + b_2 = 0, \qquad a = \pm b, \qquad t_1 \pm t_2 = 0.$$

Taking the upper sign (the work for the lower sign is similar), we get

$$a = b = \phi(u - v), \qquad t = \psi(u - v).$$

Combining this with (2) we have

$$\epsilon^t = (1/c\sqrt{?}) \cosh c(u - v), \qquad \log \theta = c(1 + i)w.$$

Using (4) we find that the refracting surface is the screw-surface

$$2c^2x = -\sinh c(u - v) \sin c(u + v), \quad 2c^2y = \sinh c(u - v) \cos c(u + v),$$
$$2cz = u + v.$$

Suppose now that the refracting surface is a sheet of the limiting surface and is not a plane.

We have $4eg = (f + f')^2$. Taking the surface in the form (1) we find that the surface is developable. Then taking it in the form of §2 (2) we obtain the congruence at the end of §2.

Take the refracting surface as plane and use §2 (1). Let the developables of all the refracted congruences meet the plane in an orthogonal net.

Then by §1 (9) $E_0f - F_0e + F_0g - G_0f' = 0$ giving

$$a_2 = b_1, \qquad a^2b_1 - aba_1 + abb_2 - b^2a_2 = 0.$$

Therefore the congruences are normal. Putting $a = \psi_1$, $b = \psi_2$ we get (3) again and can find the congruences as before.

Similarly, if the developables of all refracted congruences meet the refracting plane in a net such that the curves of the net through any point are equally inclined to a fixed direction (taken as $y = 0$), we have $E_0g + F_0(f - f') - G_0e = 0$ from §1 (9) leading to $a = \psi_2$, $b = \psi_1$ and to (3) again.

Now let the principal ruled surfaces meet the refracting plane in the same orthogonal net of straight lines for all the refracted congruences.

Using §2(1) and taking the net as $dx\,dy = 0$, we must have $\mathcal{F} = 0, f + f' = 0$ for all the refracted congruences. These lead to

$$a_2 + b_1 = 0, \qquad a_1a_2 + b_1b_2 = 0, \qquad (ab_1 - ba_1)(ab_2 - ba_2) = 0$$

whence either a is constant and b is a function of v, or $a = c\,\epsilon^{ku} \cos kv$ and $b = c\,\epsilon^{ku} \sin kv$, or else a and b have other values written down from symmetry.

More generally we can find the congruences, if the principal ruled surfaces always meet the plane in any given isometric net.

For this we take the isometric net as parametric, starting with (1) and l, m, n zero. Then $f + f' = 0$ gives $a = \epsilon^{2t}\psi_1$, $b = \epsilon^{2t}\psi_2$ and $\mathcal{F} = 0$ gives $fg + ef' = 0$, $af = be$ (or $ag = bf'$). Thence $t_2 = \{\tan^{-1}(\psi_1/\psi_2)\}_1$ and either $\psi_{11} + \psi_{22} = 0$ or

$\psi_{12} + t_1\psi_2 + t_2\psi_1 = 0$. From (2) $t = \theta(u + iv) + \Theta(u - iv)$, where θ and Θ are supposed given. The first alternative leads to

$$a = c\epsilon^{2\theta} + C\epsilon^{2\Theta}, \qquad b = i(c\epsilon^{2\theta} - C\epsilon^{2\Theta}),$$

where c and C are conjugate complex constants. If $t_2 \neq 0$ the second alternative leads to $2t_1 = -\{\log(\psi_1^2 + \psi_2^2)\}_2$ and thence to

$$a = \epsilon^t \sin t_2, \qquad b = \epsilon^t \cos t_2;$$

and similarly when $t_2 = 0$.

Consider now the case in which the developables of all the refracted congruences meet a refracting surface, which is neither plane nor minimal, in a net whose angles are bisected by a given isometric net.

With the surface in the form (1) we have from the equation (9) of §1 $E_0g + F_0(f - f') - G_0e = 0$ and thus $l = n$, $lb^2 = na^2$, $b_2 - a_1 + 2abt_1 - 2bt_2 = 0$, $(b^2 - a^2)(a_1 + b_2) = 2ab(b_1 - a_2)$. If $l = n$, $a = b$ (the work for $a = -b$ is similar), we have $a_1 = a_2, t_1 = t_2$. Hence a and b are both equal to some function of $u + v$, while also $t = \psi(u + v)$. Then (2) gives us

$$\frac{\partial q}{\partial U} = p\psi'(U), \qquad \frac{\partial p}{\partial V} = 0, \qquad pq + 2\psi''(U) = 0$$

on putting

$$l + m = p, \qquad l - m = q, \qquad u + v = U, \qquad u - v = V.$$

We then readily show that the refracting surface is a surface of revolution and that the given isometric net bisects the angles between the meridians and parallels.

As a final illustration we note the fact that, if the product of the distances of the foci of each ray from the refracting surface is unaltered by refraction, this product is the Gaussian curvature of the refracting surface at the point where it meets the ray.

4. We shall now consider briefly the case excluded heretofore, in which the ray of the congruence touches the director surface, which is therefore a sheet of the focal surface.

For such a congruence we may take the tangents to $v = \text{const.}$ as the rays of the congruence. Then $\sqrt{E}\lambda = x_1$, etc. and we find

$$e = 0, \quad E\sqrt{E}f = \{11, 2\}(EG - F^2), \quad f' = 0, \quad E\sqrt{E}g = \{12, 2\}(EG - F^2).$$

$$E(EG - F^2)\mathfrak{E} = L^2(EG - F^2) + E^2f^2,$$

$$E(EG - F^2)\mathfrak{F} = LM(EG - F^2) + E^2fg,$$

$$E(EG - F^2)\mathfrak{G} = M^2(EG - F^2) + E^2g^2.$$

The distance between the foci is

$$L\sqrt{E}/[M\{11, 2\} - L\{12, 2\}],$$

and the developables of the congruence are given by $dv(Ldu + Mdv) = 0$.

As illustrations of the use of these formulae we may quote the following results.

Any property of the congruence involving only e, f, f', g is unaltered, if the director surface is deformed carrying the rays with it. For instance, if the congruence is normal, it remains normal.

The tangent at P to any member of one family of lines of curvature of a surface forms a congruence such that the distance between the foci in the ray through P is numerically equal to the radius of geodesic curvature of the other line of curvature at P.

If the congruences formed by the two tangents to the lines of curvature at any point P of a surface have the property that the distance of P from the second focus of the congruence is the same for each tangent, the lines of curvature form an "equal-path net." The parametric net is said to be "equal-path," if $(\sqrt{E})_2 = (\sqrt{G})_1$, so that the distance between any two points of the surface is the same whatever route is taken on the surface consisting of portions of the parametric net (provided that u and v continually increase or decrease on the route).

The tangent at P is drawn to any helix of a given helicoid. If the distance between the foci of the ray through P of the congruence thus formed is constant, the helicoid is

$$x = v\cos u, \qquad y = v\sin u, \qquad z = au + c\sinh^{-1}(a/v).$$

The tangent at P to a helicoid perpendicular to the helix through P forms a normal congruence. The distance between the foci is proportional to $\sec\gamma\sec\delta$, where γ and δ are the angles which the axis of the helicoid makes with the helices and with the normal at P to the helicoid.

5. We consider lastly the problem discussed by P. Vincensini, Comptes Rendus, 190 (1930), p. 155, of the relation between a congruence and its polar reciprocal with respect to a sphere.

Let a line through the origin O meet the sphere with O as centre and unit radius at $Q(\xi, \eta, \zeta)$. For the congruence, whose director surface is the locus of $P(p\xi, p\eta, p\zeta)$, and which has the corresponding ray parallel to the tangent at Q to the curve $v = $ const. of an orthogonal parametric net on the sphere

$$\sqrt{E}e = -p_1E, \qquad \sqrt{E}f = -(p_2E + \tfrac{1}{2}pE_2), \qquad f' = 0, \qquad \sqrt{E}g = \tfrac{1}{2}pG_1.$$

$$(1) \qquad 4EG\mathfrak{E} = 4E^2G + E_2^2, \qquad 4EG\mathfrak{F} = -E_2G_1, \qquad 4EG\mathfrak{G} = G_1^2,$$

where E, F, G refer to the sphere.

Similarly for the locus of $(q\xi, q\eta, q\zeta)$ with ray parallel to the tangent to $u =$ const. at Q

$$\sqrt{G}e = \tfrac{1}{2}qE_2, \quad f = 0, \quad \sqrt{G}f' = -(q_1G + \tfrac{1}{2}qG_1), \quad \sqrt{G}g = -q_2G.$$

$$(2) \quad 4EG\mathfrak{E} = E_2^2, \quad 4EG\mathfrak{F} = -E_2G_1, \quad 4EG\mathfrak{G} = 4EG^2 + G_1^2.$$

We assume that the parametric net is not formed by coaxial great circles and their orthogonal small circles, so that E_2 and G_1 are not zero.

The two congruences are polar reciprocals of one another, if $pq = 1$.

Denoting by U some function of u only and by V some function of v, the first congruence is normal if $p^2E = U$, and is isotropic if

$$p^2 = UE, \quad U_1EG_1 + U(E_1G_1 + 4E^2G + E_2^2) = 0.$$

The product of the distances of the foci from the point P is a constant k, if $p^2 = k \log (V/G)$.

Similar relations hold for the second congruence.

We readily prove that the polar reciprocal with respect to a sphere of an isotropic congruence can be neither normal nor isotropic.

To obtain the most general normal congruence whose polar reciprocal with respect to a sphere with centre O and unit radius is also normal (Vincensini's problem) we proceed as follows. Take any parametric net on the sphere such that the sphere can be divided into small rectangles of equal area by parametric curves. Then on the sphere we may suppose

$$ds^2 = \tan\theta\, du^2 + \cot\theta\, dv^2,$$

where

$$2 + (\tan\theta)_{22} + (\cot\theta)_{11} = 0.$$

Take any point Q on the sphere and a point P on OQ such that OP is proportional to $\sqrt{(\cot\theta)}$. The line through P parallel to the tangent to $v =$ const. at Q traces out a congruence of the type required.

If OP meets a sphere with centre O in Q, and two lines through P perpendicular to OP and to each other trace out congruences (as Q varies) which are both normal, the parallel lines through Q are tangents to a fixed isometric net on the sphere.

If one of the congruences (not necessarily normal) traced out by the two lines has the locus of P as mean surface, so has the other congruence. More generally, we have the same result, if the lines are perpendicular to OP and make an angle ω with each other, while OP is a function of ω.

BEDFORD COLLEGE,
REGENTS PARK,
LONDON, ENGLAND.

ANNALS OF MATHEMATICS
Vol. 40, No. 1, January, 1939

NOTE ON A PAPER BY H. HILTON: "ON SOME PROPERTIES OF RECTILINEAR CONGRUENCES"

BY J. L. SYNGE

(Received February 3, 1938)

It is interesting to develop in tensor notation some of the ideas of Hilton's paper. This makes for compactness, reveals invariant character, and shows how his method may be extended to multidimensional cases.

Let E_N be a Euclidean space of N dimensions ($N \geqq 3$) with rectangular Cartesian coördinates x^i. Since we shall consider only such coördinates in E_N, there is no distinction between the contravariant and covariant components of tensors in E_N, and superscripts may be used uniformly. We shall use O_N to indicate an orthogonal transformation $x'^i = x'^i(x^j)$.

Let V_{N-1} be a subspace of E_N, specified by equations

$$x^i = x^i(u^\alpha), \tag{1}$$

u^α being coördinates in V_{N-1}, which is the "director-surface." Latin suffixes have the range 1, 2, $\cdots N$, Greek suffixes the range 1, 2, $\cdots N-1$, with the usual summation convention. We shall denote by T_{N-1} a general transformation $u'^\alpha = u'^\alpha(u^\beta)$.

Let us write

$$P^i_\alpha = \frac{\partial x^i}{\partial u^\alpha}; \tag{2}$$

this is a vector for O_N and a covariant vector for T_{N-1}. The line-element of V_{N-1} is

$$ds^2 = g_{\alpha\beta}du^\alpha du^\beta, \qquad g_{\alpha\beta} = P^i_\alpha P^i_\beta; \tag{3}$$

$g_{\alpha\beta}$ is the fundamental tensor of V_{N-1}, and it and its conjugate $g^{\alpha\beta}$ will be used for lowering and raising suffixes of tensors in V_{N-1} and for covariant differentiation, which will be indicated by a comma. We find for the Christoffel symbol of the first kind

$$[\alpha\beta, \gamma] = P^i_\gamma \partial P^i_\alpha/\partial u^\beta. \tag{4}$$

Hence, by (3),

$$P^i_{\alpha,\beta}P^i_\gamma = [\alpha\beta, \gamma] - \left\{ \begin{matrix} \delta \\ \alpha\beta \end{matrix} \right\} P^i_\delta P^i_\gamma = 0. \tag{5}$$

Let X^i be the unit normal vector to V_{N-1}, satisfying

$$X^i P^i_\alpha = 0, \qquad X^i X^i = 1. \tag{6}$$

Let C be a rectilinear congruence in E_N with unit vector λ^i. C may be defined by a field of unit vectors λ^i in E_N over V_{N-1}, that is, by assigning functions $\lambda^i(u^\alpha)$, satisfying the identity

$$\lambda^i\lambda^i = 1. \tag{7}$$

Hilton's plan is to define C by (7) and functions $a_\alpha(u^\beta)$ (his a, b), related to λ^i by

$$a_\alpha = \lambda^i P^i_\alpha. \tag{8}$$

Obviously a_α are invariants for O_N and components of a covariant vector for T_{N-1}. To use these quantities a_α, we require formulae giving λ^i explicitly in terms of a_α and functions of u^α which are known when the director-surface V_{N-1} is assigned. To do this, we resolve λ^i into a component along X^i and a component tangent to V_{N-1}, writing

$$\lambda^i = AX^i + t^i, \tag{9}$$

where A is an invariant and t^i a vector tangent to V_{N-1}. Now $P^i_\alpha(\alpha = 1, \cdots N-1)$ are $N-1$ vectors in E_N tangent to V_{N-1}, and hence we may write

$$t^i = b^\beta P^i_\beta, \tag{10}$$

b^β being a contravariant vector for T_{N-1}. Substituting this in (9) and multiplying by P^i_α, we get by (8), (6), (3)

$$a_\alpha = b^\beta P^i_\beta P^i_\alpha = g_{\alpha\beta}b^\beta = b_\alpha, \tag{11}$$

so that

$$t^i = a^\alpha P^i_\alpha, \qquad t^2 = t^i t^i = g_{\alpha\beta}a^\alpha a^\beta = a^2. \tag{12}$$

Since $X^i t^i = 0$, we get, on squaring and adding (9),

$$1 = A^2 + a^2, \qquad A = (1 - a^2)^{\frac{1}{2}}. \tag{13}$$

Thus the required explicit formula for λ^i *is*

$$\lambda^i = X^i(1 - a^2)^{\frac{1}{2}} + a^\alpha P^i_\alpha. \tag{14}$$

We are to note that in this formula X^i, P^i_α are determined by the director-surface alone, a, a^α by the congruence.

If ϕ denotes the angle between λ^i and X^i, we obtain, by multiplication of (14) by X^i and use of (6),

$$\cos\phi = X^i\lambda^i = (1 - a^2)^{\frac{1}{2}}, \qquad a = \sin\phi. \tag{15}$$

To obtain the condition for normality of C, we note that generally a point in E_N lying on the ray through x^i on V_{N-1} at a distance s from V_{N-1}, measured backwards, has coördinates

$$\xi^i = x^i - s\lambda^i, \tag{16}$$

ξ^i being thus a function of s and u^α, the surface-coördinates of x^i. C is a normal congruence if, and only if, s can be assigned as a function of u^α so that

$$\lambda^i \frac{\partial \xi^i}{\partial u^\alpha} = 0, \tag{17}$$

identically in u^β. But from (16), (8) and (7)

$$\lambda^i \frac{\partial \xi^i}{\partial u^\alpha} = \lambda^i \left(P^i_\alpha - s \frac{\partial \lambda^i}{\partial u^\alpha} - \lambda^i \frac{\partial s}{\partial u^\alpha} \right) = a_\alpha - \frac{\partial s}{\partial u^\alpha}. \tag{18}$$

Thus the necessary and sufficient condition that C be a normal congruence is that a_α should be a gradient,

$$a_\alpha = \partial s / \partial u^\alpha, \tag{19}$$

or, equivalently,

$$a_{\alpha,\beta} - a_{\beta,\alpha} = 0. \tag{20}$$

It is interesting to note that $s(u^\alpha)$ is Hamilton's characteristic function for a normal congruence, namely, the distance measured to a variable point u^α on the director-surface along the ray from a given normal surface. Thus the method of Hilton for skew congruences is a natural generalization of Hamilton's method, the gradient of the characteristic function being replaced by an arbitrary covariant vector field a_α in the director-surface.

Let us introduce the following notation for the coefficients of the second fundamental form of V_{N-1} and of the fundamental forms of the congruence:

$$\begin{cases} G_{\alpha\beta} = X^i_{,\alpha} P^i_\beta = -X^i P^i_{\beta,\alpha} = G_{\beta\alpha}, \\ e_{\alpha\beta} = \lambda^i_{,\alpha} P^i_\beta, \qquad E_{\alpha\beta} = \lambda^i_{,\alpha} \lambda^i_{,\beta}. \end{cases} \tag{21}$$

These are of course invariants for O_N and covariant tensors for T_{N-1}. To evaluate the latter two in terms of a_α we have merely to substitute for λ^i from (14), which gives

$$\lambda^i_{,\alpha} = X^i_{,\alpha}(1 - a^2)^{\frac{1}{2}} + X^i \frac{\partial}{\partial u^\alpha}(1 - a^2)^{\frac{1}{2}} + a^\gamma_{,\alpha} P^i_\gamma + a^\gamma P^i_{\gamma,\alpha}. \tag{22}$$

Multiplication by P^i_β gives, on account of (6) and (5),

$$e_{\alpha\beta} = G_{\alpha\beta}(1 - a^2)^{\frac{1}{2}} + a^\gamma_{,\alpha} P^i_\gamma P^i_\beta, \tag{23}$$

or, by (3),

$$e_{\alpha\beta} = a_{\beta,\alpha} + G_{\alpha\beta}(1 - a^2)^{\frac{1}{2}}. \tag{24}$$

Remembering (15), we see that these are the same as Hilton's (7), generalized to N dimensions. From (20) we obtain as conditions of normality $e_{\alpha\beta} = e_{\beta\alpha}$.

We could similarly evaluate $E_{\alpha\beta}$, but the simplicity of the expressions (8) given by Hilton appears to belong essentially to the case $N = 3$.

To deal with refraction in E_N at a V_{N-1} separating a region of refractive index

μ' from one of index μ, we assume Fermat's principle and take V_{N-1} as director-surface. The optical length of a broken line joining ξ'^i in the first medium to ξ^i in the second medium through a point x^i on V_{N-1} is

$$L = \mu' r' + \mu r, \tag{25}$$

$$r'^2 = (\xi'^i - x^i)(\xi'^i - x^i), \qquad r^2 = (\xi^i - x^i)(\xi^i - x^i).$$

The variation for a displacement of x^i on V_{N-1} is

$$dL = \mu' r'^{-1}(x^i - \xi'^i)dx^i - \mu r^{-1}(\xi^i - x^i)dx^i, \qquad dx^i = P^i_\alpha du^\alpha. \tag{26}$$

But the unit vectors along the first and second parts of the ray are

$$\lambda'^i = (x^i - \xi'^i)/r', \qquad \lambda^i = (\xi^i - x^i)/r, \tag{27}$$

and so

$$dL = (\mu' \lambda'^i P^i_\alpha - \mu \lambda^i P^i_\alpha) du^\alpha. \tag{28}$$

For a refracted ray, this is to vanish for arbitrary du^α, and so the law of refraction is, by (8),

$$\mu' a'_\alpha = \mu a_\alpha, \tag{29}$$

the natural generalization of Hilton's result.

UNIVERSITY OF TORONTO,
TORONTO, ONT., CANADA.

ANNALS OF MATHEMATICS
Vol. 40, No. 1, January, 1939

THE BEHAVIOR OF A FUNCTION ON ITS CRITICAL SET[1]

BY ANTHONY P. MORSE

(Received February 23, 1938)

1. Introduction

Referring to the next section for any unfamiliar notation or definition, let us consider the following statement.

If $m \geqq 0$, $n \geqq 1$, R is an open subset of E_n, and f is a function on R to E_1 of class C^m then f transforms its critical set into a set of (linear) measure zero.

H. Whitney[2] has shown the statement false if $m = n - 1 \geqq 1$; M. Morse and A. Sard, in an unpublished paper, have shown, on the other hand, that it is true providing m is the greatest integer in $n + (n-3)^2/16$. This last result contains as a special case the fact that the statement is true in case $m = n = 1, 2, 3, 4, 5, 6$. Thus it is natural to ask: Is the statement true in case $m = n$? Theorem 4.3 of the present paper answers this question in the affirmative. As a matter of fact the answer is yes under the less restrictive assumption that f be of class $C^{(m-1)}$ providing each partial derivative of order $(m-1)$ is totally differentiable. We give here no proof of this last statement; of the two proofs with which we are familiar one involves a fairly comprehensive modification of Section 3 as well as an additional lemma while the other involves some properties of an $(n-1)$ dimensional measure of sets contained in E_n.

An obvious corollary of Theorem 4.3 is Theorem 4.4 which states that a function on an open subset of E_n to E_1 which is of class C^n is constant on any connected subset of its critical set.

Theorems 4.3 and 4.4 are fairly direct consequences of Theorem 4.1 which is itself closely related to Theorem 3.7. Theorems 3.7 and 4.1 really form the kernel of the paper. We believe Theorem 4.1 in particular is susceptible of wider application than herein given.

2. Notations and conventions

The letters m, n, ν, j, k with or without sub- or superscripts will be permitted to assume integer values only. For $n \geqq 1$ by E_n we shall mean Euclidean space of n dimensions. If $x \in E_n$ then $| x |$ denotes the distance of x from the origin. If $A \subset E_n$ then $| A |$ denotes the (n dimensional) outer Lebesgue measure of A. By a sphere of E_n (of radius $r > 0$ and center at x') we mean an open sphere of E_n—that is the set $E_x[| x - x' | < r]$.

---

[1] Presented to the American Mathematical Society, Sept. 1938.

[2] Duke Mathematical Journal, vol. 1 (1935), page 514.

If $n \geqq 1$, $n' \geqq 1$, R is an open subset of E_n, ψ is a continuous function on R to $E_{n'}$ with

$$(\psi_1(x), \cdots, \psi_{n'}(x)) = \psi(x) \qquad (x \in R)$$

and if for each $j = 1, 2, \cdots, n'$ all partial derivatives of ψ_j through order m are continuous on R then ψ is said to be of class C^m.

If f is on $R \subset E_n$ to E_1 then for $j = 1, 2, \cdots n$ the function $D_j f$ is defined by

$$D_j f(x_1, \cdots, x_n) = \lim_{h \to 0} \frac{f(x_1, \cdots, x_j + h, \cdots, x_n) - f(x_1, \cdots, x_j, \cdots, x_n)}{h}$$

$$(x_1, \cdots, x_n) \in R;$$

furthermore a point x' for which

$$\sum_{j=1}^{n} [D_j f(x')]^2 = 0$$

will be called a *critical point* of f; the set of critical points of f will herein be called the *critical set* of f.

3. Functions whose zeros include a given set

In order to study the behavior of a function f on its critical set B, (the function f being on an open subset R of E_n to E_1 and of class C^m) it is natural, since the zeros of $D_1 f$, $D_2 f$, $\cdots$ $D_n f$ include B, first to characterize, in the neighborhood of B, the behavior of those functions which are on R to E_1, of class $C^{(m-1)}$, and which vanish on B.

We begin with

3.1 Definition. *For each positive integer n, K_n is the family of functions defined by: $\psi \in K_n$ if n' exists with $1 \leqq n' \leqq n$ such that ψ is on a sphere S of $E_{n'}$, is of class C^1, and enjoys the property,*

$$| \psi(x'') - \psi(x') | \geqq | x'' - x' | \qquad (x' \in S, x'' \in S).$$

The proofs of the next two lemmas are fairly straightforward.

3.2 Lemma. *Let $m \geqq 0$, $n \geqq 1$, R be an open subset of E_n, f be on R to E_1 and of class C^1; let $\psi \in K_n$. If $\psi(x') \in R$ and if*

$$\lim_{x \to x'} \frac{D_j f\{\psi(x)\}}{|x - x'|^m} = 0 \qquad (j = 1, 2, \cdots n)$$

then

$$\lim_{x \to x'} \frac{f\{\psi(x)\} - f\{\psi(x')\}}{|x - x'|^{(m+1)}} = 0.$$

Proof. Let

$$(\psi_1(x), \cdots, \psi_n(x)) = \psi(x)$$

for x in the domain of ψ; let $E_{n'}$ be the Euclidean space containing this domain. Let

$$\epsilon > 0, \qquad M = 1 + \sum_{j=1}^{n} \sum_{\nu=1}^{n'} [D_\nu \psi_j(x')]^2$$

and observe the existence of $\delta > 0$ such that $0 < |x - x'| < \delta$ implies

$$(1) \qquad \sum_{j=1}^{n} \frac{|D_j f\{\psi(x)\}|}{|x - x'|^m} < \frac{\epsilon}{n^{1/2} M^{1/2}}, \qquad \sum_{j=1}^{n} \sum_{\nu=1}^{n'} [D_\nu \psi_j(x)]^2 \leqq M.$$

Now let x'' be any point for which $0 < |x'' - x'| < \delta$; let $(x_1', \cdots, x_{n'}') = x'$; let $(x_1'', \cdots x_{n'}'') = x''$ and define F by

$$F(t) = f\{\psi[x' + (x'' - x')t]\} \qquad (0 \leqq t \leqq 1)$$

Thus, for $0 < t < 1$,

$$\begin{aligned} |F'(t)| &= \left| \sum_{j=1}^{n} D_j f\{\psi[x' + (x'' - x')t]\} \sum_{\nu=1}^{n'} (x_\nu'' - x_\nu') D_\nu \psi_j[x' + (x'' - x')t] \right| \\ &\leqq \left\{ \sum_{j=1}^{n} |D_j f\{\psi[x' + (x'' - x')t]\}| \right\} \\ &\qquad \cdot \sum_{j=1}^{n} \sum_{\nu=1}^{n'} |x_\nu'' - x_\nu'| \cdot |D_\nu \psi_j[x' + (x'' - x')t]| \\ &\leqq \left\{ \sum_{j=1}^{n} |D_j f\{\psi(x' + (x'' - x')t\}| \right\} \\ &\qquad \cdot \left\{ \sum_{j=1}^{n} \sum_{\nu=1}^{n'} (x_\nu'' - x_\nu')^2 \right\}^{1/2} \cdot \left\{ \sum_{j=1}^{n} \sum_{\nu=1}^{n'} [D_\nu \psi_j[x' + (x'' - x')t]]^2 \right\}^{1/2} \\ &\leqq \left\{ \sum_{j=1}^{n} |D_j f\{\psi[x' + (x'' - x')t]\}| \right\} \cdot n^{1/2} \cdot M^{1/2} \cdot |x'' - x'| \end{aligned}$$

the next to the last step being a consequence of Schwarz's inequality. Hence by the law of the mean t_1 exists, with $0 < t_1 < 1$, such that

$$\begin{aligned} \frac{|f\{\psi(x'')\} - f\{\psi(x')\}|}{|x'' - x'|^{(m+1)}} &= \frac{F(1) - F(0)}{|x'' - x'|^{m+1}} = \frac{F'(t_1)}{|x'' - x'|^{(m+1)}} \\ &\leqq \left\{ \sum_{j=1}^{n} \frac{|D_j f\{\psi[x' + (x'' - x')t_1]\}|}{|x'' - x'|^m} \right\} \cdot n^{1/2} \cdot M^{1/2} \\ &\leqq \left\{ \sum_{j=1}^{n} \frac{|D_j f\{\psi[x' + (x'' - x')t_1]\}|}{|t_1 x'' - t_1 x'|^m} \right\} \cdot n^{1/2} \cdot M^{1/2} < \epsilon, \end{aligned}$$

the last step being a consequence of (1).

The proof is complete.

3.3 LEMMA. *Let $m \geqq 1$, $n \geqq 2$; let R be an open subset of E_n and G be a function on R to E_1 of class C^m. If $y' \in Z = \underset{y}{E}[G(y) = 0]$ and if $\sum_1^n [D_j G(y')]^2 > 0$ then there exists a sphere S of E_n with y' as center and a function $\Psi \in K_n$ of class*

C^m whose domain is a sphere of E_{n-1} and whose range is included in R and includes SZ.

PROOF. We may assume without loss of generality that $D_nG(y') = 0$. Let $(y_1', \cdots, y_n') = y'$. Then by implicit function theory there exists a sphere S of E_n with y' as center and a function ϕ of class C^m whose range is contained in E_1 and whose domain is a sphere S' of E_{n-1} with $(y_1', \cdots y_{n-1}')$ as center such that

$$G(y_1, \cdots, y_{n-1}, \phi(y_1, \cdots y_{n-1})) = 0 \quad ((y_1, \cdots, y_{n-1}) \in S')$$

and such that $(y_1, \cdots, y_n) \in SZ$ implies $y_n = \phi(y_1, \cdots y_{n-1})$. Defining Ψ by $\Psi(y_1, \cdots y_{n-1}) = (y_1, \cdots, y_{n-1}, \phi(y_1, \cdots, y_{n-1}))$ $((y_1, \cdots y_{n-1}) \in S')$ as in the remainder of the proof is clear.

To clarify subsequent reasoning and to avoid considerable tedious repetition we introduce the symbolism $\langle m, n, A, B, R\rangle$ as an equivalent of the following statement:

$m \geqq 0$, $n \geqq 1$, R is an open subset of E_n and $A \subset B \subset R$; there exist sets $A_0, A_1, A_2, \cdots$ and functions $\psi_1, \psi_2, \psi_3, \cdots$ having the following properties: $A = A_0 + A_1 + A_2 + \cdots$ with A_0 countable; for each $\nu = 1, 2, 3, \cdots$ the set A_ν is bounded and condensed,[3] the function $\psi_\nu \in K_n$, its range is included in R and includes A_ν, and $\psi_\nu(x') \in A_\nu$ implies

$$\lim_{x \to x'} \frac{f\{\psi_\nu(x)\}}{|x - x'|^m} = 0$$

for each function f which is on R to E_1, of class C^m, and which vanishes on B.

Almost obvious is the following

3.4 LEMMA. *If $\langle m, n, A_j, B, R\rangle$ for $j = 1, 2, 3, \cdots$ then $\langle m, n, \sum_1^\infty A_j, B, R\rangle$*

From this and the Borel-Lebesgue covering theorem follows

3.5. LEMMA. *If $A \subset E_n$ and if corresponding to each point $y' \in A$ there is a sphere S of E_n with y' as center for which $\langle m, n, SA, B, R\rangle$ then $\langle m, n, A, B, R\rangle$.*

3.6 THEOREM. *If $m \geqq 0$, $n \geqq 1$, R is an open subset of E_n and $B \subset R$ then $\langle m, n, B, B, R\rangle$.*

PROOF. In case $n = 1$, let B^* be the set of condensation points of B; let $B_0 = B - BB^*$; for $\nu = 1, 2, 3, \cdots$ let B_ν be the intersection of BB^* with the open interval $E_x[-\nu < x < \nu]$. Observe that $B = B_0 + B_1 + B_2 + \cdots$ with B_0 countable and that for each $\nu > 0$, $x' \in B_\nu$ implies $f'(x') = f''(x') = \cdots = f^m(x') = 0$ and hence

$$\lim_{x \to x'} \frac{f(x)}{|x - x'|^m} = 0$$

for each function f which is on R to E_1, of class C^m, and which vanishes on B.

In case $m = 0$ (and $n \geqq 1$) the proof is even easier.

---

[3] We shall say a set is *condensed* if it is contained in its set of condensation points.

The theorem thus being clearly true in case $m + n = 1$ we propose to prove it by induction in $m + n$; that is we verify the following

STATEMENT 1. *If N is an integer $\geqq 1$ such that the theorem is true in case $m + n = N$, and if $m' \geqq 0$, $n' \geqq 1$, $m' + n' = N + 1$, R' is an open subset of E_n and $B' \subset R'$ then $\langle m', n', B', B', R' \rangle$.*

Since the truth of $\langle m', n', B', B', R' \rangle$ has already been established in case $m' = 0$ or in case $n' = 1$ we henceforth assume $m' \geqq 1$, $n' \geqq 2$ and let α be the set defined by: $y \in \alpha$ if $y \in B'$ and

$$\sum_{j=1}^{n'} [D_j f(y)]^2 = 0$$

for *each* function f which is on R' to E_1, of class $C^{m'}$, and which vanishes on B'. Defining $\beta = B' - \alpha$ we divide the remainder of the proof of Statement 1 into four parts.

PART I. $\langle m', n', \alpha, B', R' \rangle$.

From the hypotheses of Statement 1 we know that $\langle m' - 1, n', \alpha, \alpha, R' \rangle$ which implies the existence of sets $\alpha_0, \alpha_1, \alpha_2, \cdots$ and functions $\psi_1, \psi_2, \psi_3, \cdots$ with the following properties: $\alpha = \alpha_0 + \alpha_1 + \alpha_2 + \cdots$ with α_0 countable; for each $\nu = 1, 2, 3, \cdots$ the set α_ν is bounded and condensed, the function $\psi_\nu \in K_n$, its range is included in R' and includes α_ν, and $\psi_\nu(x') \in \alpha_\nu$ implies

$$\lim_{x \to x'} \frac{g\{\psi_\nu(x)\}}{|x - x'|^{(m'-1)}} = 0$$

for each function g which is on R' to E_1, of class $C^{(m'-1)}$ and which vanishes on α. Letting f be any function which is on R' to E_1, of class $C^{m'}$, and which vanishes on B' we observe that the definition of α assures us that for $j = 1, 2, \cdots, n'$ the function $D_j f$ is one which is on R' to E_1, of class $C^{(m'-1)}$, and which vanishes on α, and hence (identifying $D_j f$ with g above) has the property that $\psi_\nu(x') \in \alpha_\nu$ implies

$$\lim_{x \to x'} \frac{D_j f\{\psi_\nu(x)\}}{|x - x'|^{m'-1}} = 0,$$

which in turn implies (see 3.2)

$$\lim_{x \to x'} \frac{f\{\psi_\nu(x)\}}{|x - x'|^{m'}} = 0.$$

PART II. *If $y' \in \beta = B' - \alpha$ then there exists a sphere S of $E_{n'}$ with y' as center such that $\langle m', n', S\beta, B', R' \rangle$.*

By the definition of β there exists a function G which is on R' to E_1, of class $C^{m'}$, which vanishes on B', and which has the property

$$\sum_{j=1}^{n'} [D_j G(y')]^2 > 0.$$

Hence by 3.3 there exists a sphere S of $E_{n'}$ with y' as center and a function $\Psi \in K_{n'}$ of class $C^{m'}$ whose domain is a sphere S' of $E_{n'-1}$ and whose range is included in R' and includes $S\beta$. Univalence (one-to-oneness) of Ψ follows from the relation

$$| \Psi(x'') - \Psi(x') | \geqq | x'' - x' | \qquad (x' \in S', x'' \in S'),$$

which is a consequence of Ψ being an element of $K_{n'}$. Let γ be the subset of S' which Ψ maps into $S\beta$. Since $\langle m', n' - 1, \gamma, \gamma, S' \rangle$ by the hypotheses of Statement 1, we conclude the existence of sets $\gamma_0, \gamma_1, \gamma_2, \cdots$ and functions $\phi_1, \phi_2, \phi_3, \cdots$ having the following properties: $\gamma = \gamma_0 + \gamma_1 + \gamma_2 + \cdots$ with γ_0 countable; for each $\nu = 1, 2, 3, \cdots$ the set γ_ν is bounded and condensed, the function $\phi_\nu \in K_{n'-1}$, its range is included in S' and includes γ_ν and $\phi_\nu(x') \in \gamma_\nu$ implies

$$\lim_{x \to x'} \frac{g\{\phi_\nu(x)\}}{|x - x'|^{m'}} = 0$$

for each function g which is on S' to E_1, of class $C^{m'}$, and which vanishes on γ.

Let B_0 be the set into which Ψ maps γ_0; for each $\nu = 1, 2, 3, \cdots$ let β_ν be the set into which Ψ maps γ_ν and let ψ_ν be the function defined by

$$\psi_\nu(x) = \Psi\{\phi_\nu(x)\} \qquad (x \in \text{the domain of } \phi_\nu).$$

Now for $\nu = 1, 2, 3, \cdots$ it is clear that ψ_ν is of class C', that

$$| \psi_\nu(x'') - \psi_\nu(x') | \geqq | \phi_\nu(x'') - \phi_\nu(x') | \geqq | x'' - x' |$$

for x' and x'' in the domain of ϕ_ν and that ψ_ν possesses the other properties requisite that ψ_ν be an element of $K_{n'}$. Thus it becomes apparent that $S\beta = \beta_0 + \beta_1 + \beta_2 + \cdots$ with β_0 countable, that for each $\nu = 1, 2, 3, \cdots$ the set β_ν is condensed[4] and bounded, the function $\psi_\nu \in K_{n'}$, its range is included in R' and includes β_ν.

We complete the proof of Part II by letting f be any function which is on R' to E_1, of class $C^{m'}$, and which vanishes on B', observing that the function F defined by

$$F(x) = f\{\Psi(x)\} \qquad (x \in S')$$

is one which is on S' to E_1, of class $C^{m'}$, and which vanishes on γ, and noticing that, by identifying F with the g which occurs at the end of the first paragraph of this part, it follows that $\phi_\nu(x') \in \gamma_\nu$ implies

$$\lim_{x \to x'} \frac{f\{\psi_\nu(x)\}}{|x - x'|^{m'}} = \lim_{x \to x'} \frac{F\{\phi_\nu(x)\}}{|x - x'|^{m'}} = 0;$$

[4] By virtue of the univalence of Ψ and the fact that γ_ν is condensed.

that is $\psi_\nu(x') \in \beta_\nu$ implies

$$\lim_{x \to x'} \frac{f\{\psi_\nu(x)\}}{|x - x'|^{m'}} = 0.$$

PART III. $\langle m', n', \beta, B', R' \rangle$.

This is an immediate consequence of Part II and Lemma 3.5.

PART IV. $\langle m', n', B', B', R' \rangle$.

Since $\alpha + \beta = B'$, Parts I and III combine with Lemma 3.4 to yield $\langle m', n', B', B', R' \rangle$.

The proof of Theorem 3.6 is now complete. Translating its symbolic conclusion we obtain

3.7 THEOREM. *If $m \geqq 0$, $n \geqq 1$, R is an open subset of E_n and $B \subset R$ then there exist sets $B_0, B_1, B_2, \cdots$ and functions $\psi_1, \psi_2, \psi_3, \cdots$ having the following properties: $B = B_0 + B_1 + B_2 + \cdots$ with B_0 countable; for each $\nu = 1, 2, 3, \cdots$ the set B_ν is bounded and condensed, the function $\psi_\nu \in K_{n'}$, its range is included in R and includes B_ν, and $\psi_\nu(x') \in B_\nu$ implies*

$$\lim_{x \to x'} \frac{f\{\psi_\nu(x)\}}{|x - x'|^m} = 0$$

for each function f which is on R to E_1, of class C^m, and which vanishes on B.

4. The Behavior of a Function on its Critical Set

In view of 3.7 and 3.2 we have

4.1 THEOREM. *If $m \geqq 1$, $n \geqq 1$, R is an open subset of E_n, and $B \subset R$ then there exist sets $B_0, B_1, B_2, \cdots$ and functions $\psi_1, \psi_2, \psi_3, \cdots$ having the following properties: $B = B_0 + B_1 + B_2 + \cdots$ with B_0 countable; for each $\nu = 1, 2, 3, \cdots$ the set B_ν is bounded and condensed, the function $\psi_\nu \in K_n$, its range is included in R and includes B_ν, and $\psi_\nu(x') \in B_\nu$ implies*

$$\lim_{x \to x'} \frac{f\{\psi_\nu(x)\} - f\{\psi_\nu(x')\}}{|x - x'|^m} = 0$$

for each function f which is on R to E_1, of class C^m, and whose critical set includes B.

From the last property ascribed to a function $\in K_n$ by Definition 3.1 it is clear from 4.1 that we have

4.2 THEOREM. *If $m \geqq 1$, $n \geqq 1$, R is an open subset of E_n, and $B \subset R$ then there exist sets $B_0, B_1, B_2, \cdots$ having the following properties: $B = B_0 + B_1 + B_2 + \cdots$ with B_0 countable; for each $\nu = 1, 2, 3, \cdots$ the set B_ν is bounded and condensed and $y' \in B_\nu$ implies*

$$\lim_{\substack{y \to y' \\ y \in B_\nu}} \frac{f(y) - f(y')}{|y - y'|^m} = 0$$

for each function f which is on R to E_1, of class C^m, and whose critical set includes B.

We now come to

4.3 THEOREM. *If $n \geqq 1$, R is an open subset of E_n, and f is a function on R to E_1 of class C^n then f transforms its critical set into a set of measure zero.*

PROOF. For each set $\alpha \subset R$ we shall denote by $f^*(\alpha)$ the set into which f transforms α. Letting B be the critical set of f it is then our purpose to prove $|f^*(B)| = 0$. From 4.2 follows the existence of sets $B_0, B_1, B_2, \cdots$ having the following properties: $B = B_0 + B_1 + B_2 + \cdots$ with B_0 countable; for each $\nu = 1, 2, 3, \cdots$ the set B_ν is bounded and condensed and $y' \in B_\nu$ implies

$$\lim_{\substack{y \to y' \\ y \in B_\nu}} \frac{f(y) - f(y')}{|y - y'|^n} = 0.$$

Now let $\nu_0 \geqq 1$, $A = B_{\nu_0}$, $\epsilon > 0$ and for $k = 1, 2, 3, \cdots$ define A_k by: $y' \in A_k$ if $y' \in A$ and if

$$|f(y) - f(y')| \leqq \frac{\epsilon |y - y'|^n}{n^n(|A| + 1)}$$

for each point $y \in A$ which is distant $\leqq 1/k$ from y'. (Clearly $|A| < \infty$.) For each $k = 1, 2, 3, \cdots$ let $I_{1,k}, I_{2,k}, I_{3,k}, \cdots$ be open cubes of E_n, each of diameter $\leqq 1/k$, which have the property that $A \subset I_{1,k} + I_{2,k} + I_{3,k} + \cdots$ and $|I_{1,k}| + |I_{2,k}| + |I_{3,k}| + \cdots \leqq |A| + 1$. Thus $y' \in I_{j,k} A_k$ and $y'' \in I_{j,k} A_k$ together imply

$$|f(x'') - f(y')| \leqq \frac{\epsilon |y'' - y'|^n}{n^n(|A| + 1)} \leqq \frac{\epsilon n^{n/2} |I_{j,k}|}{n^n(|A| + 1)} \leqq \frac{\epsilon}{|A| + 1} |I_{j,k}|,$$

which in turn implies

$$|f^*(I_{j,k} A_k)| \leqq \frac{\epsilon}{|A| + 1} \cdot |I_{j,k}|.$$

Whence for $k = 1, 2, 3, \cdots$ it follows that

$$|f^*(A_k)| = \left| \sum_{j=1}^{\infty} f^*(I_{j,k} A_k) \right| \leqq \sum_{j=1}^{\infty} |f^*(I_{j,k} A_k)| \leqq \frac{\epsilon}{|A| + 1} \sum_{j=1}^{\infty} |I_{j,k}| \leqq \epsilon,$$

and since (see definition of A_k) the relations

$$A = A_1 + A_2 + A_3 + \cdots, \qquad A_k \subset A_{k+1} \qquad (k = 1, 2, 3 \cdots),$$

imply

$$f^*(A) = f^*(A_1) + f^*(A_2) + f^*(A_3) + \cdots, \quad f^*(A_k) \subset f^*(A_{k+1}) \quad (k = 1, 2, 3 \cdots),$$

we conclude[5]

$$|f^*(A)| = \left| \sum_{k=1}^{\infty} f^*(A_k) \right| = \lim_{k \to \infty} |f^*(A_k)| \leqq \epsilon.$$

The arbitrariness of ϵ implies $|f^*(A)| = |f^*(B_{\nu_0})| = 0$.

---

[5] There is no necessity for knowing that $f^*(A_k)$ is measurable. See, for example, Titchmarsh, *The Theory of Functions*, Oxford, 1932, page 329.

Consequently $| f^*(B_\nu) | = 0$ for $\nu = 1, 2, 3, \cdots$ and since $f^*(B_0)$ is countable it is clear that

$$| f^*(B) | = \left| \sum_{\nu=0}^{\infty} f^*(B_\nu) \right| \leqq \sum_{\nu=0}^{\infty} | f^*(B_\nu) | = 0.$$

In conclusion we prove

4.4 THEOREM. *If $n \geqq 1$, R is an open subset of E_n, and f is a function on R to E_1 of class C^n then f is constant on every connected subset of its critical set.*

PROOF. Let B be any connected subset of the critical set of f. Since f is continuous we conclude in the notation of 4.3 that $f^*(B)$ is connected and hence an interval. But by 4.3, $| f^*(B) | = 0$ so that $f^*(B)$ is either void or has precisely one element.

The proof is complete

THE INSTITUTE FOR ADVANCED STUDY.

ANNALS OF MATHEMATICS
Vol. 40, No. 1, January, 1939

THE PARSEVAL THEOREM OF THE CAUCHY SERIES AND THE INNER PRODUCTS OF CERTAIN HILBERT SPACES

BY TOSIO KITAGAWA

(Received April 20, 1938)

1. Introduction

On the realizations of an abstract Euclidean space,[1] that is, a complex linear metric space in which an inner product is defined, we have only a few concrete examples of the inner products. Thus in the space $L^2(a, b)$, the familiar form of the inner product $(f, g) = \int_a^b f(t)\overline{g(t)}\,dt$ is perhaps the sole one which appears in the literature.[2] It is due to this form that, in the study[3] of symmetric transformations associated with the ordinary differential operators, the boundary conditions are confined to those which depend solely on the end values of the concurrence-functions.[4] If the boundary condition is another form,[5] then it will be found that that form of the inner product is not convenient. The first object of this paper is to introduce a perhaps new form of the inner product

---

[1] For its precise formulation, see H. Löwig: *Komplexe euklidische Räume von beliebiger endlicher oder transfiniter Dimensionszahl*, Acta Litt. Ac Scien., 7 (1934), pp. 1–33.

[2] As another concrete example, we may mention

$$(f, g) = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} f(t)\overline{g(t)}\,dt,$$

in the space of all the functions which are almost periodic in the sense of Bohr. Cf. Fr. Rellich: *Spektraltheorie in nichtseparabeln Räumen*, Math., Ann., 110 (1935), pp. 342–356.

[3] Cf. M. H. Stone: *Linear transformations in Hilbert space and their applications to analysis*. Am. Math. Coll. Volume XV, especially Chapter X, §2 and 3. In the following we shall quote this treatise as [S].

[4] See [S], Theorem 10.7 on p. 428 and Theorem 10.11 on p. 458.

[5] The general boundary-value problems of the linear differential operations were discussed by many authors. For example, see J. D. Tamarkin: *Some general problems of the theory of ordinary linear differential equations and expansion of an arbitrary function in series of fundamental functions*. Math. Zeitschr. 27 (1926), pp. 1–54.

J. D. Tamarkin discussed a linear differential equation

$$y^{(n)} + P_1(x, \rho)y^{(n-1)} + \cdots + P_n(x, \rho)y = 0,$$

with $P_i(x, \rho) = \rho i \sum_{j=0}^{i} \rho^{-j} p_{i,j}(x)$ and under n boundary conditions of the following type:

$$L_i(y) = \sum_{k=1}^{n} \sum_{\nu=1}^{\infty} \gamma_{i,k}^{(\nu)}\, y^{(k-1)}(t_{i,k}^{(\nu)}) + \int_a^b y^{(n-1)}(t)\, d\gamma_{i,k}(t).$$

which is suitable for the application of abstract theory of Hilbert space to the formal differential operator $Df(x) = f'(x)/i$ under the boundary condition $l_\xi\{f(\xi)\} \equiv \int_{a-0}^{b+0} f(\xi)d\varphi(\xi) = 0$, where $\varphi(\xi)$ is a function of bounded variation on the closed finite interval (a, b). For the sake of brevity we may and shall assume $a = 0$ and $b = 2\pi$, without any loss of generality. The method of introducing a new inner product is as follows. First with each function $f(x)$ belonging to $L^2(0, 2\pi)$ we shall associate its Cauchy series with respect to l.[6] Secondly we shall prolong the domain of definition of $f(x)$ to the interval $(-2\pi, 0)$, and thus obtain a function $f_1(x)$ which belongs to $L^2(-2\pi, 2\pi)$ and coincides with $f(x)$ in the original interval $(0, 2\pi)$. After these prolongations, we s[1] all define

$$(1) \qquad (f, g) \equiv l_\xi\left\{\int_0^\xi f_1(\eta)\overline{g_1(\eta - \xi)}\, d\eta\right\} \equiv \int_{-0}^{2\pi+0}\left\{\int_0^\xi f_1(\eta)\overline{g_1(\eta - \xi)}\, d\eta\right\}d\varphi(\xi).$$

This scalar function of elements f and g can be expressed by the coefficients of Cauchy expansions of f and g. Here we shall appeal to the results of Paley-Wiener and Levinson on the non-harmonic Fourier series.[7] The relation may be called the Parseval relation of Cauchy series with respect to the linear functional l. We shall show that, under certain restrictions[8] on the generating function $G(\lambda) \equiv l_\xi\{e^{\lambda\xi}\}$, the new bilinear form (f, g) is a positive definite Hermitian form. Thus we can form a Hilbert space by choosing the form (f, g) as an inner product in $L^2(0, 2\pi)$. We shall have incidentally improved the superfluous assumptions under which we established the Parseval relation previously.[9]

The second purpose is to establish an operational calculus of the linear translatable operation[10] on the function space $\mathfrak{A}_\varphi$ which consists of all the functions which (i) are defined over $(-\infty, \infty)$ and L^2-integrable in any finite interval of reals and for which (ii) $\Lambda f(x) \equiv \int_{-0}^{2\pi+0} f(x+t)\, d\varphi(t) = 0$, for all x in $-\infty < x < \infty$, except perhaps for an x-set of zero measure.

We shall show that (i) with each element $f(x)$ of $\mathfrak{A}_\varphi$ we may associate its essential Cauchy series; (ii) for each pair of elements $f(x)$ and $g(x)$ belonging to $\mathfrak{A}_\varphi$, we can define their essential inner product; (iii) the complex linear space $\mathfrak{A}_\varphi$ constitutes a Hilbert space by introducing these essential inner products. If

---

[6] The precise definition will be given in §2.

[7] Paley-Wiener: *Fourier transforms in the complex domain*. Am. Math. Coll. XIX, Chap. VII, pp. 100–116.

N. Levinson: *On Non-harmonic Fourier Series*, Annals Math. 37 (1936), pp. 917–936. We shall quote these as [P. W.] and [L] respectively.

[8] That is under the conditions 1°, 2° and 3° which will be given afterwards.

[9] T. Kitagawa, *On the theory of linear translatable functional equation and Cauchy's series*, Jap. Journ. Math., XIII, (1937), pp. 233–322; see especially Chapter III, §12, *A Certain Generalization of Parseval's Theorem*, pp. 285–290. We shall quote this as [T]. Cf. also T. Kitagawa: *The multiplication-theorem of Cauchy series*, Proc. Imp. Acad. Tokyo (1938).

[10] For its definition, see §5.

we add the condition 4°[11] on $\varphi(\xi)$, then the differential operator $f'(x)/i$ is a self-adjoint operator with simple spectrum, and consequently the operational calculus of all the bounded linear translatable operations can be treated by the spectral theory of abstract Hilbert space.

2. Introduction of Cauchy series with respect to the given linear functional

We shall consider a contour-integral[12]

$$S_C(x,f) \equiv \frac{1}{2\pi i}\oint_C \frac{e^{\lambda x}}{G(\lambda)} l_\xi\left\{e^{\lambda\xi}\int_0^\xi e^{-\lambda\eta}f(\eta)\,d\eta\right\}d\lambda. \tag{2}$$

If we choose a sequence of contours $\{C_r\}$ such that C_r is contained in the interior of the domain enclosed by C_{r+1}, and C_r diverges to the whole plane as $r\to\infty$,[13] we shall obtain a series of the exponential sum, which we shall call the Cauchy series of $f(x)$ with respect to l. In what follows we shall confine ourselves to the case which satisfies the following

CONDITION 1°. *All the zero-points of $G(\lambda)$ are simple and imaginary; we shall arrange them in $\{i\lambda_n\}$ in the following manner:*

$$\cdots < \lambda_{-n} < \lambda_{-n+1} < \lambda_{-n+2} < \cdots < \lambda_{-1} < \lambda_0 < \lambda_1 < \cdots < \lambda_n < \lambda_{n+1} < \cdots$$

We shall first prove a lemma.

LEMMA 1.[14] *Under condition 1°, let us define the set of functions $\{h_n(\eta)\}$ by*

$$h_n(\eta) = \frac{e^{-i\lambda_n\eta}}{G'(i\lambda_n)}\int_{\eta-0}^{2\pi+0} e^{i\lambda_n\xi}\,d\varphi(\xi),\quad (n=0,\pm1,\pm2,\cdots). \tag{3}$$

Then $\{e^{-i\lambda_n\eta}\}$ and $\{h_n(\eta)\}$ form a biorthogonal function-system, that is,

$$\int_0^{2\pi} e^{i\lambda_m\eta}h_n(\eta)\,d\eta = \begin{cases}1, & [m=n],\\ 0, & [m\neq n].\end{cases} \tag{4}$$

---

[11] The condition 4° will be given in §5.

[12] In [T], we considered the Cauchy series with respect to the linear translatable operations. The corresponding form is

$$Sc_r(x, x_0; f) \equiv \frac{1}{2\pi i}\oint_{C_r}\frac{e^{\lambda x}}{G(\lambda)}\Lambda_\xi\left\{e^{\lambda\xi}\int_{x_0}^{x_0+\xi} e^{-\lambda\eta}f(\eta)\,d\eta\right\}d\lambda. \tag{1.01}$$

See [T], Chapter I, §1. *Transformation of Contour-integral*, (1.01) on p. 245 and Remark 1°, (1.20) on p. 251. Thus (1.01) is a special case of the integral (2).

[13] For the precise meaning see [T], p. 240.

[14] Under certain different assumptions, we have previously proved the corresponding fact in [T], Chapter III, §14, *Integral-representation of the Biorthogonal Functions*, where it should be read:

$$h_n(\eta) = \frac{e^{-i\alpha\eta}}{G'(i\lambda_n)}\int_{\eta-0}^{b+0} e^{i\lambda_n\xi}\,d\varphi(\xi), \tag{14.04}$$

Further we have

$$\frac{h_n(\eta+0)+h_n(\eta-0)}{2}=\frac{1}{G'(i\lambda_n)}\operatorname*{l.i.m.}_{T\to\infty}\int_{-iT}^{iT}\frac{G(z)e^{\eta z}}{i\lambda_n-z}dz. \tag{5}$$

PROOF: Let $f(x)$ be an element of $L^2(0, 2\pi)$ with the Cauchy expansion

$$f(x)\backsim\sum_{-\infty}^{\infty}a_k e^{i\lambda_k x}. \tag{6}$$

Then by the contour-integral (1), we have

$$a_n=\frac{1}{G'(i\lambda_n)}\int_{-0}^{2\pi+0}\left\{\int_0^{\xi}e^{i\lambda_m(\xi-\eta)}f(\eta)\,d\eta\right\}d\varphi(\xi). \tag{7}$$

By the interchange of the order of the Lebesgue-Stieltjes integration, we have

$$a_n=\frac{1}{G'(i\lambda_n)}\int_0^{2\pi}f(\eta)e^{-i\lambda_n\eta}\left(\int_{\eta-0}^{2\pi+0}e^{i\lambda_n\xi}\,d\varphi(\xi)\right)d\eta. \tag{8}$$

and as is readily seen,

$$\frac{1}{G'(i\lambda_n)}\int_{-0}^{2\pi+0}\left\{\int_0^{\xi}e^{i\lambda_n(\xi-\eta)}e^{i\lambda_m\eta}\,d\eta\right\}d\varphi(\xi)=\begin{cases}1, & [m=n],\\ 0, & [m\pm n].\end{cases} \tag{9}$$

The proof of (4) can be obtained by a Laplace integral.[15] q.e.d.

We shall introduce

CONDITION 2°.[16] *There is a constant D such that* $|\lambda_n-n|\leqq D<\frac{1}{\pi^2}$, $(-\infty<n<\infty)$. This leads to

LEMMA 2. *Let $f(x)$ be any function belonging to $L^2(0, 2\pi)$. Then under conditions 1°-2°, we have the following*:

(i) *The series*

$$\sum_{-\infty}^{\infty}\left\{\frac{e^{inx}}{2\pi}\int_0^{2\pi}f(\xi)e^{-in\xi}\,d\xi-e^{i\lambda_n x}\int_0^{2\pi}f(\xi)h_n(\xi)\,d\xi\right\} \tag{10}$$

converges uniformly to zero over any interval $(-\epsilon, 2\pi+\epsilon)$ and over any such interval the convergence and summability properties of the series

$$\sum_{-\infty}^{\infty}e^{i\lambda_n x}\int_0^{2\pi}f(\xi)h_n(\xi)\,d\xi\equiv\sum_{-\infty}^{\infty}a_n e^{i\lambda_n x} \tag{11}$$

---

[15] See [T], Chapter III, §14, pp. 291–292.

[16] This condition is that of Paley-Wiener. The best possible constant is that of Levinson which is 1/4. Since we shall make use of the relation (iii) in Lemma 2, we shall be content with the less sharp one.

are the same as those of the ordinary Fourier series

$$\sum_{-\infty}^{\infty} \frac{e^{inx}}{2\pi} \int_0^{2\pi} f(\xi) e^{-in\xi} d\xi, \tag{12}$$

(ii) $$(1 - \pi D^{\frac{1}{2}})^2 \sum_{n=-\infty}^{\infty} |a_n|^2 dx \leqq \int_0^{2\pi} |f(x)|^2 dx \leqq (1 + \pi D^{\frac{1}{2}})^2 \sum_{-\infty}^{\infty} |a_n|^2.$$

(iii) *The series*

$$\sum_{-\infty}^{\infty} a_n e^{-i\lambda_n x} \tag{13}$$

converges in the mean[17] *to a function belonging to* $L^2(-2\pi, 0)$.

PROOF: To prove (i)–(ii), it suffices to remark that, by Lemma 1 and conditions 1°–2°, the Cauchy expansion of $f(x)$ with respect to the linear functional l is identical with the non-harmonic Fourier series with respect to the set $\{e^{i\lambda_n x}\}$ in the terminology of Paley-Wiener and Levinson (iii) follows immediately by combining (i) and (ii).[18]

In the following lines we shall always prolong the function $f(x)$ belonging to $L^2(0, 2\pi)$ to the function $f_1(x)$ which belongs to $L^2(-2\pi, 2\pi)$ and which coincides with $f(x)$ in $(0, 2\pi)$ by (iii) in Lemma 2.

3. The Parseval relation of Cauchy series

LEMMA 3. *Let $f(x)$ and $g(x)$ be any two functions belonging to $L^2(0, 2\pi)$. Let their Cauchy series with respect to l be*

$$f(x) \sim \sum_{-\infty}^{\infty} a_n e^{i\lambda_n x}, \tag{6}$$

$$g(x) \sim \sum_{-\infty}^{\infty} b_n e^{i\lambda_n x}, \tag{14}$$

respectively. We shall assume that they are prolonged to the functions $f_1(x)$ and $g_1(x)$ belonging to $L^2(-2\pi, 2\pi)$ respectively by the method given in §2. Then

$$(1) \qquad (f, g) \equiv l_\xi \left\{ \int_0^\xi f_1(\eta) \overline{g_1(\eta - \xi)}\, d\eta \right\} = \lim_{n \to \infty} \sum_{-n}^{n} a_k \bar{b}_k G'(i\lambda_k).$$

---

[17] The convergence in the mean will be understood throughout this paper in the following sense: $\{f_n(x)\}$ $(n = 1, 2, \cdots)$ is said to be convergent to $f(x)$ in the mean over an interval (a, b), if

$$\lim_{n \to \infty} \int_a^b |f_n(t) - f(t)|^2 dt = 0.$$

[18] Cf. [L], Lemma 4 in p. 927.

This follows from Theorem XXXVII and the inequality (30.07) in [P. W.]. See [P. W.] (iii) p. 100 and p. 108.

PROOF: Let us put $g_{(n)}(x) = \sum_{k=-n}^{n} b_k e^{i\lambda_k x}$. Then $\int_0^\xi f_1(\eta)\overline{g(\eta - \xi)}\,d\eta$ is a continuous function of ξ in the interval $(-2\pi, 2\pi)$. Direct computation gives, in view of Lemma 1,

$$l_\xi\left\{\int_0^\xi f_1(\eta)\overline{g_{(n)}(\eta - \xi)}\,d\eta\right\} = \sum_{k=-n}^{n} a_k \overline{b_k} G'(i\lambda_k). \tag{15}$$

On the other hand we have

$$\left|\int_0^\xi f_1(\eta)\overline{g_1(\eta - \xi)}\,d\eta - \int_0^\xi f_1(\eta)\overline{g_{(n)}(\eta - \xi)}\,d\eta\right| \leqq \sqrt{\int_{-2\pi}^{2\pi} |f_1(\eta)|^2\,d\eta}\sqrt{\int_{-2\pi}^{2\pi} |g_1(\eta) - g_{(n)}(\eta)|^2\,d\eta}. \tag{16}$$

Since $\{g_{(n)}(\xi)\}$ tends to $g(\xi)$ in the mean over $(-2\pi, 2\pi)$, this shows that the sequence of continuous functions $\left\{\int_0^\xi f_1(\eta)g_{(n)}(\eta - \xi)\,d\eta\right\}$ tends uniformly to $\int_0^\xi f_1(\eta)\overline{g_1(\eta - \xi)}\,d\eta$ on the interval $(0, 2\pi)$. Consequently it follows that

$$\lim_{n\to\infty} l_\xi\left\{\int_0^\xi f_1(\eta)\overline{g_{(n)}(\eta - \xi)}\,d\eta\right\} = l_\xi\left\{\int_0^\xi f_1(\eta)\overline{g_1(\eta - \xi)}\,d\eta\right\}.$$

Combining this with (16), we reach the result to be proved.

LEMMA 4. *The bilinear form (f, g) possesses the following properties*: (1°) $(af, g) = a(f, g)$, *for any given complex number* a; (2°) $(f + g, h) = (f, h) + (g, h)$; (3°) $(f, g) = \overline{(g, f)}$; (4°) *If* $G'(i\lambda_n) > 0$ *for* $n = 0, \pm 1, \pm 2, \cdots$, *then* $f \neq 0$ *and* $f = 0$ *implies that* $(f, f) > 0$ *and* $(f, f) = 0$ *respectively.*

PROOF: Obvious from Lemma 3.

Lemma 3 implies that to any given $f(x)$ belonging to $L^2(0, 2\pi)$ there corresponds a sequence $\{a_k\}$ $(k = 0, \pm 1, \pm 2, \cdots)$ such that $\sum_{k=-\infty}^{\infty} |a_k|^2 G'(i\lambda_k)$ is convergent. We shall show the converse under the hypothesis of the following

CONDITION 3°. *There is a pair of positive constants* α *and* β *such that* $\beta \geqq G'(i\lambda_n) \geqq \alpha > 0$, *for* $n = 0, \pm 1, \pm 2, \cdots$.

We shall prove

LEMMA 5. *Let us assume that* $G(\lambda)$ *satisfies conditions* 1°, 2° *and* 3°. *Then for any given sequence* $\{a_k\}$ $(k = 0, \pm 1, \pm 2, \cdots)$ *subject to the condition that* $\sum_{-\infty}^{\infty} |a_k|^2 G'(i\lambda_k) < \infty$, *there corresponds a uniquely determined function* $f(x)$ *belonging to* $L^2(0, 2\pi)$ *whose Cauchy expansion is given by*: $f(x) \sim \sum_{-\infty}^{\infty} a_k e^{i\lambda_k x}$.

PROOF: Condition 3° implies that $\sum_{-\infty}^{\infty} |a_k|^2 < \infty$. If we put $f_{(N)}(x) \equiv \sum_{-N}^{N} a_k e^{i\lambda_k x}$, then we have, in view of Lemma 2 (iii),

$$(17) \quad \int_0^{2\pi} |f_{(N_1)}(x) - f_{(N_2)}(x)|^2 dx \leqq (1 + \pi D^{\frac{1}{2}})^2 \left\{ \sum_{-N_2}^{-N_1} |a_n|^2 + \sum_{N_1}^{N_2} |a_n|^2 \right\}.$$

Therefore we know that $\{f_{(n)}(x)\}$ converges to a function $f(x)$ in the mean on the interval $(0, 2\pi)$ and that $f(x) \sim \sum_{-\infty}^{\infty} a_k e^{i\lambda_k x}$.

Summing up the results, we reach

THEOREM I. *Let $\varphi(\xi)$ be a function of bounded variations on $(0, 2\pi)$. Let us assume that the generating function $G(\lambda) \equiv l_\xi\{e^{\lambda\xi}\} \equiv \int_{-0}^{2\pi+0} e^{\lambda\xi} d\varphi(\xi)$ is subject to the conditions 1°, 2° and 3°. Then the complex linear space of all functions which are defined on $(0, 2\pi)$ and L^2-integrable, constitutes a Hilbert space, denoted by $L^2_\varphi(0, 2\pi)$ if we define the inner product of any two elements f and g by (15) given in Lemma 3:* $\left(\text{Then } \|f\|^2 = l_\xi\left\{\int_0^\xi f_1(\eta)\overline{f_1(\eta - \xi)}\, d\eta\right\}\right)$

This may be recognized as the Parseval relation of Cauchy series.[19]

4. The symmetric transformations associated with differentiation

Let D^* be the class of all functions expressible in the form $f(x) = c + \int_0^x g(\xi)d\xi$ where c is an arbitrary constant and $g(\xi)$ is an arbitrary function in $L^2(0, 2\pi)$. Let D be the class of all functions expressible in the form $f(x) = \int_0^x g(\xi)d\xi$ where $g(\xi)$ is an arbitrary function in $L^2(0, 2\pi)$ such that $l_\xi\left\{\int_0^\xi g(\eta)\, d\eta\right\} = 0$.

---

[19] In the Fourier series, we may put

$$\varphi(\xi) = \begin{cases} -\frac{1}{2}, & [\xi = 0], \\ 0, & [0 < \xi < 2\pi], \\ \frac{1}{2}, & [\xi = 2\pi]. \end{cases}$$

Incidentally, it should be remarked that in my previous formulation we considered

$$\int_{-\pi+0}^{\pi+0} \left(\int_0^t g(\eta)\overline{f(t-\eta)}\, d\eta\right) d\varphi(t),$$

that is,

$$\int_{-0}^{2\pi+0} \left(\int_0^t g(\eta)\overline{f(t-\eta)}\, d\eta\right) d\varphi(t).$$

But this is not convenient to ascertain that $(f, f) > 0$, although we may assume condition 4°, which we shall employ in the next §5.

Furthermore, let $D(\theta)$ be the class of all functions such that $l_\xi\{e^{i\theta\xi}f(\xi)\} = 0$ for a value of θ in $(0, 2\pi)$. We shall denote by H, $H(\theta)$ and T the transformations with the domains D, $D(\theta)$ and D^* resp., each of which takes an arbitrary function $f(x)$ in its domain into the function $f'(x)/i$. We shall observe

THEOREM II. *In the Hilbert space $L^2_\varphi(0, 2\pi)$ defined in §3, the operators H, $H(\theta)$ and T possess the following properties*: (i) *H is a closed linear symmetric transformation with the adjoint $H^* \equiv T$ and the deficiency index* (1, 1); *and the transformation $H(\theta)$ is a self-adjoint extension of H*; (ii) *The transformation $H(\theta)$ has a simple point spectrum, consisting of the characteristic values $\lambda_n - \theta (n = 0, \pm 1, \pm 2, \cdots)$ with the corresponding characteristic functions $\{e^{i(\lambda_n - \theta)x}\}$.*

This is a generalization of Theorem 10.7 in Stone's[20] treatise to which our Theorem II corresponds word by word. The proof[21] runs parallel to that of Stone's Theorem, if we notice that by integration by parts we have

$$l_\xi\left\{\int_0^\xi \frac{f'(\eta)}{i}\,\overline{g(\eta-\xi)}\,d\eta\right\} = \frac{\overline{g(0)}}{i}\,l_\xi\{f(\xi)\} - \frac{f(0)}{i}\,l_\xi\{\overline{g(-\xi)}\} - l_\xi\left\{\int_0^\xi f(\eta)\,\frac{\overline{g'(\eta-\xi)}}{i}\,d\eta\right\}. \tag{18}$$

5. The operational calculus of the bounded linear translatable operators on the space $\mathfrak{A}_\varphi$

We shall now consider the complex linear space defined in the Introduction. The fact that each element $f(x)$ is a solution of the functional equation, that is, in $-\infty < x < \infty$,

$$\Lambda f(x) \equiv l_\xi\{f(x+\xi)\} \equiv \int_{-0}^{2\pi+0} f(x+\xi)\,d\varphi(\xi) = 0, \tag{19}$$

except perhaps for an x-set of zero measure, implies the following:

(i) Lemma 1 tells us that for any given x in $(-\infty, \infty)$, the Cauchy series defined by the sequence of the contour-integral

$$\frac{1}{2\pi i}\oint_C \frac{e^{\lambda x}}{\overline{G(\lambda)}}\,l_\xi\left\{e^{\lambda\xi}\int_0^\xi e^{-\lambda\eta}f(x_0+\eta)\,d\eta\right\}d\lambda \tag{20}$$

converges in the mean to $f(x)$ on the interval $(x_0, x_0 + 2\pi)$. $f(x)$ being a solution of (19), we can verify that the Cauchy series generated by (20) is inde-

---

[20] See [S] p. 428.

[21] It is to be noted that, in the proof of Theorem II, we shall make use of the following rather obvious Lemma: *To any given constants α and β, there is a function $g(x)$ belonging to $L^2(0, 2\pi)$ such that $g(0) = \alpha$ and $l_\xi\{g(-\xi)\} = \beta$.*

To prove this, we shall choose a positive constant k and a complex number v such that $l_\xi\{(-\xi)^k\} \neq 0$ and $l_\xi\{e^{-v\xi}\} \neq 0$. Then the following function satisfies the required property:

$$g(x) = \alpha e^{-vx} + \frac{\beta - \alpha l_\xi\{e^{-v\xi}\}}{l_\xi\{(-\xi)^k\}}\,(-x)^k.$$

pendent of x_0 in $-\infty < x_0 < \infty$, except perhaps for an x_0-set of zero measure; almost everywhere as the essential Cauchy expansion of $f(x)$.

(ii) For any two given $f(x)$ and $g(x)$, both of which belong to $\mathfrak{A}_\varphi$, the bilinear form

$$l_\xi\left\{\int_0^\xi f(\sigma+\eta)\overline{g(\tau+\eta-\xi)}\,d\eta\right\} \tag{21}$$

is independent of both of σ and τ in $-\infty < \tau, \sigma < \infty$, except perhaps for certain σ-set and τ-set, each of which is of zero measure; this common value is called the essential inner-product of f and g in $\mathfrak{A}_\varphi$. It can be readily seen that if the essential Cauchy expansions of f and g are given by $f \backsim \sum a_n e^{i\lambda_n x}$ and $g \backsim \sum b_n e^{i\lambda_n x}$, then the essential inner-product of them amounts to $\sum_{-\infty}^{\infty} a_k \overline{b_k} G'(i\lambda_k)$, which we shall denote by (f, g).

Thus in view of Theorem II it can be readily seen that, under conditions 1°–3°, the complex linear space $\mathfrak{A}_\varphi$ constitutes a Hilbert space by introducing the essential inner product defined in (ii). From now on we shall add

CONDITION 4°. There is a constant h such that $\varphi(\xi) = h\varphi(\xi - 2\pi)$,[22] $0 \leqq \xi < 2\pi$. Then it follows that for any function $f(x)$ belonging to $\mathfrak{A}_\varphi$,

$$\int_{-0}^{2\pi+0} f(x-\xi)\,d\varphi(\xi) = 0 \tag{22}$$

almost everywhere in $(-\infty, \infty)$. As in (18), we have

$$l_\xi\left\{\int_0^\xi \frac{f'(\sigma+\eta)}{i}\overline{g(\tau+\eta-\xi)}\,d\eta\right\} = \frac{g(\tau)}{i}l_\xi\{f(\sigma+\xi)\} - \frac{f(\sigma)}{i}l_\xi\{\overline{g(\tau-\xi)}\}$$
$$- l_\xi\left\{\int_0^\xi f(\sigma+\eta)\frac{\overline{g'(\tau+\eta-\xi)}}{i}\,d\eta\right\},$$

provided that $f(x)$ and $g(x)$ are absolutely continuous in any finite interval. This relation, in view of (19), (21) and (22), amounts to $(f'/i, g) = (f, g'/i)$.

On account of that we find immediately

THEOREM III. *Let D be the class of all functions $f(x)$ in $\mathfrak{A}_\varphi$ such that $f(x)$ is absolutely continuous in any finite range and has a derivative $f'(x)$ belonging to $\mathfrak{A}_\varphi$. Let D be transformation which takes an arbitrarily function $f(x)$ in D into the function $f'(x)/i$. Then under the conditions 1°–4°, the complex linear space $\mathfrak{A}_\varphi$ can be recognized as the Hilbert space with our essential product, and D is a self-adjoint operator with the simple spectrum $\{\lambda_n\}(n = 0, \pm 1, \pm 2, \cdots)$ and the corresponding characteristic functions $\{i^{i\lambda_n x}\}$.*

An operator A in $\mathfrak{A}_\varphi$ is said to be a bounded linear translatable operator, if, besides its linearity and boundedness, A possesses the following property: if both of $f(x)$ and $Af(x)$ belong to the domain of D, then $ADf(x) = DAf(x)$.

[22] In the case of the ordinary Fourier Series, we may assume the condition 4° is satisfied, by defining $\varphi(\xi)$ as in Footnote 19.

We can now apply the spectral theory of abstract Hilbert space to the operational calculus of bounded linear translatable operators. The resolution of the identity corresponding to D, $E(\lambda)$, is as follows: for f with the expansion $f \backsim \sum_{-\infty}^{\infty} a_n e^{i\lambda_n x}$ in $\mathfrak{A}_\varphi$, $(E(\lambda)f, f) = \sum_{\lambda_n \leq \lambda} a_n \overline{a_n} G'(i\lambda_n)$; especially when all the coefficients $a_n \pm 0$, then the closed linear manifold determined by the set of elements $E(\lambda)f$, $-\infty < \lambda < \infty$, coincides with whole space. Due to the fact that the self-adjoint operator D has a simple spectrum, the necessary and sufficient condition that a linear bounded operator should be translatable is that it is a function[23] of D, and we can establish the operational calculus of all the bounded linear translatable operators in $\mathfrak{A}_\varphi$.

Appendix I

Here we wish to point out the problem of deciding whether the conditions 1°, and 2° would imply that there is a function of bounded variation such that $G(\lambda) = \int_{-0}^{2\pi+0} e^{\lambda t}\, d\varphi(t)$. If the problem were solved, the connection between the Non-harmonic Fourier series and the Cauchy series would become transparent.

Appendix II

The inner product introduced in this paper is peculiarly favored by the fact that the operation considered are the differentiation and the translatable operators. For the boundary-value problems of the differential operators of higher order with variable coefficients[24] our inner product is not suitable. It should be noted that for the ordinary inner product multiplication possesses the simple property that

$$\int_a^b r(t) f(t) \overline{g(t)}\, dt = \int_a^b f(t) \overline{\overline{r(t)} g(t)}\, dt.$$

The corresponding simple relation for our inner product is difficult to establish.

MATHEMATICAL INSTITUTE, ÔSAKA IMPERIAL UNIVERSITY.

[23] For the precise description, see [S], Theorem 8.1 on p. 300. Also cf. [S] Chapter VI, *The operational calculus*.

[24] See [S], Chapter X, §3.

ANNALS OF MATHEMATICS
Vol. 40, No. 1, January, 1939

THE NUMBER OF GROUPS WHICH INVOLVE A GIVEN NUMBER OF UNITY CONGRUENCES, AND APPLICATIONS

BY D. T. SIGLEY

(Received April 3, 1938)

1. Introduction

The number of distinct abstract groups of an order which is the product of distinct prime factors and which involves less than four unity congruences has been determined by Miller.[1] The same author[2] has determined all the orders for which there exists a given number k of groups, for all values of $k < 6$. Miller[3] also has proposed the problem of determining whether or not there exists an order for which there is an arbitrary number k of finite abstract groups. All of these problems are somewhat related and additional results will be given for each of the problems. The main objective of this paper is the determination of the number of distinct abstract groups of squarefree order in which there are precisely four unity congruences. The results of this determination, table I, are used as an aid in the determination of the orders for which there exist exactly six groups, table II, and in proving theorem 4. Miller's results have been included in the tables to make them complete.

2. The number of prime factors involved in four unity congruences

Let the order g of a finite abstract group be the product of distinct prime factors. We shall be interested in the number of groups of order g when the number of unity congruences involved in g is four. The number of prime factors h which involve an arbitrary number m of unity congruences may be any number between k and $2m$, where k is the largest integer such that $C(k, 2) \leqq m$. For $m = 4$, it is found that $h = 4, 5, 6, 7$, or 8. Hence, to determine the number of groups of an order which is the product of distinct prime factors which involve four unity congruences, we may consider separately the cases in which the number of distinct prime factors involving the congruences is an integer between four and eight inclusive. The order g may be multiplied by any number of distinct prime factors which are also prime to the factors of g and which do not increase the number of unity congruences in the resulting order without affecting the number of groups of this resulting order.

If the order g involves four unity congruences, and the primes are arranged in order of magnitude, the four unity congruences may be selected in 15 ways.

[1] G. A. Miller, Amer. Jour. Math., vol. LV (1933), pp. 22-28.
[2] G. A. Miller, Proc. Nat. Acad. Sci., vol. 18, two papers.
[3] G. A. Miller, *Collected Works*, vol. I (1935), p. 93.

The number of groups in each of these cases has been determined recently.[4] When the number of distinct prime factors of g is five, the unity congruences may be selected in 135 ways, if the primes are arranged in the order of descending magnitude. If the number of distinct prime factors of g exceeds five, the order g may be factored into the product of two relatively prime factors of the form of g such that no prime factor of one of the factors is congruent to unity with respect to a prime factor of the other factor. The direct product theorem[1] applies, and the number of groups of order g of these forms may be written from the results of the determinations in the earlier cases.

3. The number of groups of a squarefree order which involves precisely four unity congruences

From the preceding section it is seen that the only new results needed for the enumeration of the groups of an order g which is the product of distinct prime factors and which involves four unity congruences are those for which the number of prime factors of g which involve the congruence relations is five. Each of these 135 cases have been considered separately with only the results recorded in the table. The method is the same as that used in the determination of the groups of order $pqrs$[4] and is outlined below.

Throughout this paper the letters p, q, r, s, t, u, v, and w, with or without subscripts, represent distinct prime numbers. The unity congruence relation $p \equiv 1$, mod q, shall be represented by $p(q)$. By the *number of groups* of order g we shall understand the *number of distinct finite abstract groups* of order g.

The groups of order g are solvable. We shall enumerate them according to the order of the central. This order may be unity, or any divisor of g greater than unity, except a divisor of the form $g/(\text{a prime})$. If the central is of order g, G is abelian, and there is always one and only one group. If the central quotient group has an order of the form pq, then G is the direct product of a non-abelian group of an order of the form pq and a cyclic group of order g/pq. Hence the number of groups of this type is equal to the number of unity congruences in g. Next, let the order of the central quotient group be of the form pqr. A group G of this type will be the direct product of a group of order pqr in which the central is the identity and a cyclic group of order g/pqr. If $P, Q, R, S, T, \cdots$ represent operators of orders $p, q, r, s, t, \cdots\cdots$, respectively, in G of order $g = pqrst\cdots$, we may consider G to be generated by these operators. We now select all of the possible triples of generators, whose orders are subject to the given congruence relations, as generators of the groups of orders of the form pqr. The number of groups of an order of the form pqr in which the central is the identity is $r - 1$, 0, 1, or r, according as the congruence relations are $p(r)$, $q(r)$; $p(q)$, $q(r)$; $p(q)$, $p(r)$; or $p(q)$, $p(r)$, and $q(r)$. This process may be repeated until we arrive at the case in which the central is assumed to be the identity.

[4] 'A determination of the groups of order $pqrs$,' to be published by the Amer. Jour. Math., January 1939.

If the central G is the identity we select the defining relations so that no operator of order greater than 1 is invariant. If we assume that g involves k unity congruences and assume also that the number of groups of an order of the form of g involving less than k unity congruences is known, then we may write the number of groups in which the generators fulfil less than k defining relations of the form $Q^{-1} P Q = P^{\alpha \neq 1}$. There remains only the case in which the generators fulfil the maximum number k of these defining relations. Hence, the enumeration of the groups of squarefree order involving k unity congruences may be enumerated from the results of the same problem when the number of unity congruences is less than k and the enumeration of the groups when all of the possible pairs of generators are non-commutative as allowed by the unity congruences in the order g.

As an illustration we shall enumerate the groups of order $pqrst$ subject to the congruence relations $p(q)$, $p(r)$, $q(t)$, and $r(s)$. There is *one* abelian group. There are *four* (four unity congruences) groups in which the central is of an order of the form pqr. If we consider as invariant the pairs of generators PQ, PR, PS, PT, QR, QS, QT, RS, RT, and ST, we find there is *one* group when the pair ST is invariant and no group in the other cases. There is one group when P, S, or T is the only invariant generator, and no group when Q, or R is the only invariant generator. According to the unity congruence relations we may have (1) $Q^{-1}PQ = P^{\alpha \neq 1}$, (2) $R^{-1}PR = P^{\rho \neq 1}$, (3) $T^{-1}QT = Q^{\delta \neq 1}$, and (4) $S^{-1}RS = R^{\omega \neq 1}$, or any subset of these, in a set of defining relations of G with the condition that no generator is invariant. In order that neither S nor T may be invariant (3) and (4) must be in every set of defining relations of G. In order that P is invariant, (1), or (2), or both, must be satisfied, but (3) contradicts (1) and (4) contradicts (2). There is no group of this special order in which the central is the identity. The total number of groups of this order is *nine*.

In table I, the number of groups of a squarefree order which involves k, $0 \leqq k \leqq 4$, unity congruences is given. When g involves four unity congruences, the cases are classified according to the number of prime factors involved in the unity congruences. Also in the cases when the number of factors is four or five, the prime factors in alphabetic order are arranged in the order of descending magnitude. In the remainder of the table this convention does not hold, and the sample entries will include all cases.

TABLE I

The number of groups of an order which is the product of distinct prime factors and which involves k unity congruences for values of k between 0 and 4

	Unity congruence relations	Number of groups
	No unity congruence.	
1.		1
	One unity congruence.	
1.	$p(q)$	2

Two unity congruences.[5]

1.	$p(q)\ q(r)$	3
2.	$p(q)\ p(r)$	4
3.	$p(r)\ q(r)$	$r+2$
4.	$p(q)\ r(s)$	4

Three unity congruences.[5]

1.	$p(q)\ q(r)\ s(t)$	6
2.	$p(q)\ p(r)\ s(t)$	8
3.	$p(r)\ q(r)\ s(t)$	$2r+4$
4.	$p(q)\ r(s)\ t(u)$	8
5.	$p(q)\ p(r)\ p(s)$	8
6.	$p(q)\ p(r)\ q(r)$	$r+4$
7.	$p(q)\ p(r)\ q(s)$ or $r(s)$	6
8.	$p(q)\ q(r)\ q(s)$	5
9.	$p(q)\ q(r)\ r(s)$	5
10.	$p(r)\ q(r)\ r(s)$	$r+3$
11.	$p(q)\ q(s)\ r(s)$	$s+4$
12.	$p(s)\ q(s)\ r(s)$	s^2+s+2

Four unity congruences.

Unity congruences on four prime factors

1.	$p(q)\ p(r)\ p(s)\ q(r)$	$2r+8$
2.	$q(s)$[6]	$2s+8$
3.	$r(s)$	$2s+8$
4.	$q(r)\ q(s)$	$2r+8$
5.	$r(s)$	$r+6$
6.	$q(s)\ r(s)$	$s+7$
7.	$p(s)\ q(r)\ q(s)$	$2s+6$
8.	$r(s)$	$2s+6$
9.	$q(s)\ r(s)$	s^2+2s+4
10.	$q(r)\ q(s)\ r(s)$	$s+6$
11.	$p(r)\ p(s)\ q(r)\ q(s)$	$rs+2r+2s+4$
12.	$r(s)$	$2r+s+4$
13.	$q(s)\ r(s)$	s^2+2s+4
14.	$q(r)\ q(s)\ r(s)$	$2r+s+4$
15.	$p(s)\ q(r)\ q(s)\ r(s)$	s^2+2s+4

Unity congruences on five prime factors

1.	$p(q)\ p(r)\ p(s)\ p(t)$	16
2.	$q(t)$	12
3.	$r(t)$	12
4.	$s(t)$	12
5.	$p(t)\ q(s)$	12
6.	$r(s)$	12
7.	$s(t)$	$4t+8$
8.	$q(r)\ s(t)$	$2r+8$
9.	$q(s)\ q(t)$	9

[5] The prime factors in the order are not necessarily in the order of decreasing magnitude.
[6] The remaining congruences are the same as the corresponding ones of the preceding line.

Unity congruences on five prime factors—*Continued*

10.	$r(t)$	9
11.	$s(t)$	10
12.	$q(t)\ r(s)$	9
13.	$s(t)$	$2t + 8$
14.	$r(s)\ r(t)$	10
15.	$s(t)$	10
16.	$r(t)\ s(t)$	$2t + 8$
17.	$p(s)\ p(t)\ q(r)$	12
18.	$r(s)$	$4s + 8$
19.	$r(t)$	$4t + 8$
20.	$q(r)\ q(t)$	10
21.	$r(t)$	10
22.	$s(t)$	9
23.	$q(s)\ r(t)$	$2s + 8$
24.	$q(t)\ r(s)$	$2s + 8$
25.	$r(t)$	$2t + 8$
26.	$r(s)\ r(t)$	$4s + 8$
27.	$s(t)$	$2s + 6$
28.	$r(t)\ s(t)$	$2t + 8$
29.	$p(t)\ q(r)\ q(s)$	10
30.	$r(s)$	10
31.	$s(t)$	$3t + 6$
32.	$q(s)\ r(s)$	$2s + 8$
33.	$r(t)$	$3t + 6$
34.	$q(t)\ r(s)$	$2t + 8$
35.	$r(s)\ r(t)$	$4t + 8$
36.	$s(t)$	$2t + 8$
37.	$r(t)\ s(t)$	$2t^2 + 2t + 4$
38.	$q(r)\ q(s)\ q(t)$	9
39.	$r(t)$	8
40.	$s(t)$	8
41.	$q(t)\ r(s)$	8
42.	$s(t)$	$2t + 6$
43.	$r(s)\ r(t)$	9
44.	$s(t)$	8
45.	$r(t)\ s(t)$	$2t + 6$
46.	$q(s)\ q(t)\ r(s)$	$2s + 6$
47.	$r(t)$	$2t + 6$
48.	$r(s)\ r(t)$	$2s + 8$
49.	$s(t)$	$s + 6$
50.	$r(t)\ s(t)$	$2t + 6$
51.	$q(t)\ r(s)\ r(t)$	$2t + 8$
52.	$s(t)$	$t + 7$
53.	$r(t)\ s(t)$	$t^2 + 2t + 4$
54.	$r(s)\ r(t)\ s(t)$	$2t + 8$
55.	$p(r)\ p(s)\ p(t)\ q(r)$	$4r + 8$
56.	$q(s)$	$4s + 8$
57.	$q(t)$	$4t + 8$
58.	$q(r)\ q(t)$	$4r + 8$
59.	$r(t)$	$2r + 6$
60.	$s(t)$	$3r + 6$

Unity congruences on five prime factors—*Continued*

61.	$q(s)\ q(t)$	$4s + 8$
62.	$r(t)$	$3s + 6$
63.	$s(t)$	$2s + 6$
64.	$q(t)\ r(s)$	$2s + 8$
65.	$r(t)$	$2t + 8$
66.	$s(t)$	$2t + 8$
67.	$p(t)\ q(r)\ q(s)$	$4r + 8$
68.	$r(s)$	$2r + 6$
69.	$s(t)$	$rt + 2r + 2t + 4$
70.	$q(s)\ q(t)$	$4t + 8$
71.	$r(s)$	$2s + 8$
72.	$r(t)$	$2t + 8$
73.	$s(t)$	$2t + 8$
74.	$q(t)\ r(s)$	$3t + 6$
75.	$s(t)$	$2t^2 + 2t + 4$
76.	$q(r)\ q(s)\ q(t)$	$4r + 8$
77.	$r(t)$	$2r + 6$
78.	$s(t)$	$3r + 6$
79.	$q(t)\ r(s)$	$2r + 6$
80.	$s(t)$	$rt + 2r + 2t + 4$
81.	$r(s)\ r(t)$	$r + 5$
82.	$s(t)$	$2r + 5$
83.	$r(t)\ s(t)$	$2r + t + 4$
84.	$q(s)\ q(t)\ r(s)$	$2s + 8$
85.	$r(t)$	$2t + 8$
86.	$s(t)$	$2t + 8$
87.	$r(s)\ r(t)$	$2s + 6$
88.	$s(t)$	$s + 6$
89.	$r(t)\ s(t)$	$t + 7$
90.	$q(t)\ r(s)\ r(t)$	$2t + 6$
91.	$s(t)$	$2t + 6$
92.	$r(t)\ s(t)$	$t^2 + 2t + 4$
93.	$p(s)\ p(t)\ q(s)$	$4s + 8$
94.	$q(t)$	$4s + 8$
95.	$r(s)$	$2s + 8$
96.	$r(t)$	$2t + 8$
97.	$s(t)$	$2t + 8$
98.	$q(s)\ r(s)$	$2s^2 + 2s + 4$
99.	$r(t)$	$st + 2s + 2t + 4$
100.	$q(t)\ r(s)$	$st + 2s + 2t + 4$
101.	$r(t)$	$2t^2 + 2t + 4$
102.	$q(r)\ q(s)\ q(t)$	$4s + 9$
103.	$r(t)$	$3s + 6$
104.	$s(t)$	$2s + 6$
105.	$q(t)\ r(s)$	$2s + 8$
106.	$r(t)$	$2t + 8$
107.	$s(t)$	$2t + 7$
108.	$r(s)\ r(t)$	$2s + 6$
109.	$s(t)$	$s + 6$
110.	$r(t)\ s(t)$	$t + 7$
111.	$q(s)\ q(t)\ r(s)$	$2s^2 + 2s + 4$

Unity congruences on five prime factors—*Concluded*

112.	$r(t)$	$st + 2s + 2t + 4$
113.	$r(s)\ r(t)$	$2s^2 + 2s + 4$
114.	$s(t)$	$s^2 + s + 3$
115.	$r(t)\ s(t)$	$2s + t + 4$
116.	$q(t)\ r(s)\ r(t)$	$st + 2s + 2t + 4$
117.	$s(t)$	$2s + t + 4$
118.	$r(t)\ s(t)$	$t^2 + 2t + 4$
119.	$p(t)\ q(r)\ q(s)\ q(t)$	$4t + 8$
120.	$r(s)$	$2s + 8$
121.	$r(t)$	$2t + 8$
122.	$s(t)$	$2t + 8$
123.	$q(t)\ r(s)$	$2t + 6$
124.	$s(t)$	$2t^2 + 2t + 4$
125.	$r(s)\ r(t)$	$2t + 6$
126.	$s(t)$	$2t + 6$
127.	$r(t)\ s(t)$	$t^2 + t + 4$
128.	$q(s)\ q(t)\ r(s)$	$st + 2s + 2t + 4$
129.	$r(t)$	$2t^2 + 2t + 4$
130.	$r(s)\ r(t)$	$st + 2s + 2t + 4$
131.	$s(t)$	$2s + t + 4$
132.	$r(t)\ s(t)$	$t^2 + 2t + 4$
133.	$q(t)\ r(s)\ r(t)$	$2t^2 + 2t + 4$
134.	$s(t)$	$t^2 + 2t + 4$
135.	$q(t)\ r(t)\ s(t)$	$t^3 + t^2 + t + 2.$

Unity congruences on[5] six prime factors

1.	$p(q)\ q(r)\ s(t)\ t(u)$	9
2.	$s(u)$	12
3.	$s(u)\ t(u)$	$3u + 6$
4.	$p(q)\ p(r)\ s(t)\ s(u)$	16
5.	$s(u)\ t(u)$	$4u + 8$
6.	$p(r)\ q(r)\ s(u)\ t(u)$	$ru + 2r + 2u + 4$

Unity congruences on[5] seven prime factors

1.	$p(q)\ q(r)\ s(t)\ u(v)$	12
2.	$p(q)\ p(r)\ s(t)\ u(v)$	16
3.	$p(r)\ q(r)\ s(t)\ u(v)$	$4r + 8$

Unity congruences on[5] eight prime factors

1.	$p(q)\ r(s)\ t(u)\ v(w)$	16

4. Orders for which there exist precisely six groups[7]

THEOREM 1. *If the smallest number of groups of an order g which involves α unity congruences is k, then there exist at least $k + \rho$ groups of an order gg', g' an integer, which involves $\alpha + \rho$ unity congruences.*

The proof of the theorem follows from the fact that the introduction of an

[7] Read before the American Mathematical Society. See Bull. Amer. Math. Soc., vol. XLII (1936), p. 816.

additional unity congruence in the order of a group by multiplying the order by a factor increases the number of groups by at least one.

COROLLARY. *There are at least $k + 1$ groups of an order g which involves k unity congruences.*

We shall determine all of the orders for which there are exactly six groups. Since it is possible under certain conditions to multiply an order g by a square-free factor without increasing the number of groups, and since we wish to give the general form of the orders from which six groups exist, we shall give a more general statement of Miller's direct product theorem.

THEOREM 2. *If a number g may be separated into the product of two relatively prime factors g_1 and g_2, where g_2 is the product of distinct prime factors, such that no prime factor of g_1 (or g_2) divides the order of the group of isomorphisms of g_2 (or g_1), then the number of groups of order g is the product of the number of groups of the two orders g_1 and g_2.*

This theorem may be proved as follows. Let G_1 and G_2 be two groups of orders g_1 and g_2, respectively, where g_1 and g_2 satisfy the hypotheses stated in the theorem. Since no prime factor of g_1 (or g_2) divides the order of the group of isomorphisms of G_2 (or G_1), in a group G of order $g = g_1g_2$ every operator of G_1 is commutative with every operator of G_2, and G is the direct product of G_1 and G_2. Therefore, the theorem is true.

The number of groups of order 16 is 14, and the number of groups of order p^4, p an odd prime, is 15, and hence an order g for which there exist precisely six groups is not divisible by the fourth power of a prime factor. If the highest power of a prime factor which divides g is the third, then g is not divisible by the square of another prime factor. Neither does g involve a unity congruence, and from the known[8] number of groups of an order of the form p^3q, it follows that there precisely six groups of an order g which is divisible by the cube of a prime factor if and only if, g involves no unity congruence, no prime factor of g divides $p + 1$, and exactly one prime factor of g divides $p^2 + p + 1$. The order of g is necessarily odd.

If the square of a prime factor is the highest power of a prime which divides g, then g is not divisible by the squares of more than two distinct prime factors. If g is divisible by the squares of two distinct primes, then g cannot involve a unity congruence. Accordingly, there are precisely six groups of an order of the form $p^2q^2rst \cdots$, if and only if (a) $p + 1$ or $q + 1$, but not both, is divisible by exactly one prime factor of g different from q or p^1, or (b) $p + 1$ is divisible by q, but not by q^2.[1]

If g is divisible by the square of exactly one prime factor, and by no higher power of a prime, g involves less than three unity congruences according to theorem 1. When the number of unity congruences is zero the number of groups of order g is of the form $1 + 2^{\alpha 1}$, which is not equal to six for an integral value of α. If g involves one unity congruence, there are precisely six groups of order g

[8] A. E. Western, Proc. London Math. Soc., vol. 30 (1899), pp. 209–263.

under the following conditions: the prime p whose square divides g is not involved in the unity congruence, and one prime factor of g not involved in the unity congruence divides $p + 1$. If g involves two unity congruences, the prime p whose square divides g is not involved in either congruence, nor is $p + 1$ divisible by a prime factor of g. The direct product theorem applies and there are exactly six groups of an order of the form under consideration if and only

TABLE II

Table of orders for which there is a given number of groups

Number of groups	General form of the order	Congruence relations	Sample order
1	$pqrst \cdots$	No uc[9]	15
	$pqrst \cdots$	$p(q)$	21
	$p^2qrst \cdots$	No uc, and no prime divides $p + 1$	45
3	$pqrst \cdots$	$p(q)$, $q(r)$	609
	$p^2qrst \cdots$	No uc, 1 prime divides $p + 1$	75
4	$2pq$	$p(2)$, $q(2)$	30
	$pqrst \cdots$	$p(q)$, $r(s)$	1155
	$pqrst \cdots$	$p(q)$, $p(r)$	465
	$p^2qrst \cdots$	$q(r)$, no prime divides $p + 1$	495
	$p^2qrst \cdots$	$q(p)$, not $q(p^2)$, no prime divides $p + 1$	28
	$p^2q^2rst \cdots$	No uc, no prime divides $p + 1$ or $q + 1$	1225
5	$4p$	$p = 3$, or $p(4)$	12 or 20
	$3pqrs \cdots$	$p(3)$, $q(3)$	273
	$pqrst \cdots$	$p(q)$, $q(r)$, $q(s)$	317,595
	$pqrst \cdots$	$p(q)$, $q(r)$, $r(s)$	35,931
	$2p^2$	$p(2)$	18
	$p^2qrst \cdots$	$q(p^2)$	1,332
	$p^2qrst \cdots$	No uc, 2 primes divide $p + 1$	12,615
	$p^2qrst \cdots$	$q(p)$, not $q(p^2)$, one prime divides $p + 1$	825
	$p^3qrst \cdots$	no uc, no prime divides $p + 1$, or $p^2 + p + 1$	8
6	$2pq$	$p(q)$, $p(2)$, $q(2)$	42
	$pqrst \cdots$	$p(q)$, $q(r)$, $q(s)$	7455
	$pqrst \cdots$	$p(q)$, $q(r)$, $s(t)$	33,495
	$p^2qrst \cdots$	$q(r)$, 1 prime not q or r divides $p + 1$	65,065
	$p^2qrst \cdots$	$q(r)$, $r(s)$, no prime divides $p + 1$	735,063
	$p^2q^2rs \cdots$	No uc, q but not q^2 divides $p + 1$	1,089
	$p^2q^2rs \cdots$	No uc, and one prime, not q, divides $p + 1$	29,575
	$p^3qrst \cdots$	No uc, no prime divides $p + 1$, one prime divides $p^2 + p + 1$	6517

if three of the prime factors of g different from p are such that the first is congruent to unity with respect to the second and the second is thus congruent with respect to the third.

Lastly, we consider g to be the product of distinct primes. From the results of table I and theorem 1 it follows that g involves less than four unity con-

[9] uc—unity congruence(s).

gruences. From the first part of table we find that there are precisely six groups of order $g = pqrstuv \cdots$, if and only if (a) $p(q)$, $q(r)$ $s(t)$, (b) $g = 2pq$, with $p(q)$, and (c) $p(q)$, $p(r)$, $r(s)$.

In table II we give the general form of the order, with the congruence relations involved, and an example of each order for which there exist precisely 1, 2, 3, 4, 5, or 6 finite abstract groups.

5. Orders for which a given number of groups exist

We shall now investigate the existence of an order g for which there exists an arbitrary given number k of finite abstract groups. The proof of the existence of at least one order g for every positive integer k is a problem which has not yet been solved, although all of the orders for values of k not exceeding 6 are known. A sufficient condition for the existence of an order g for a given k may be proved for a special case by the aid of the direct product theorem.

THEOREM 3. *If there exists an order g which is product of distinct prime factors for which there are precisely k groups, such that k depends only on the number of prime factors of g and the number of unity congruences involving these prime factors, then there exists an order g' for which there are precisely k^α, $\alpha > 0$, abstract groups.*

PROOF: Assuming the existence of g in the theorem, we may construct another order g_1 of the same form as g such that the following are true: (a) g_1 involves the same number of prime factors as g; (b) g_1 involves the same number of unity congruences as g, and the prime factors of g and g_1 may be arranged so that there is a $1 - 1$ correspondence between the unity congruences of g and those of g_1; and (c) no unity congruence in the product gg_1 involves prime factors from both orders g and g_1. Now, by the direct product theorem there are k^2 groups of the order gg_1. By complete induction the theorem follows.

We prove the following theorem on the existence of an order g for which there are precisely k groups, for values of k of a special form.

THEOREM 4. *There exists at least one order g for which there are precisely k abstract groups, where k is of the form $2^\alpha 3^\delta 5^\omega 7^\rho p$, where α, δ, ω, and ρ are positive integers or zeros and p is a prime of the form $q + 2$, $q + 4$, $q + 6$, $2q + r + 4$, $q^2 + 2q + 4$, or $q^2 + q + 3$, where q and r are distinct odd prime numbers.*

Since there exists an infinite number of primes among the terms of an arithmetic progression in which the common difference is prime to the first term, there exists a number g which is the product of m distinct prime factors and which involves an arbitrary combination of x possible unity congruences for $x \leqq C(m, 2)$. The order for which there is a number of groups equal to each of the prime powers 2^α, 3^δ, 5^ω, 7^ρ, and p, of the theorem, will be shown to exist by demonstrating the order.

The order for which there are precisely 2^α groups may be constructed as the product of 2α distinct prime factors p_i, $i = 1, 2, 3, \cdots 2\alpha$, for which we have the congruence relations p_{2i-1} (p_{2i}), for $i = 1, 2, \cdots, \alpha$. The order for which there are precisely 3^δ abstract groups may be constructed as the product of 3δ distinct prime factors which may be divided into δ sets of 3 each such that the

first factor in each set is congruent to unity with respect to the second factor, and the second factor is thus congruent with respect to the third factor of the set. The order for which there are precisely 5^ω groups may be constructed as the product of the third powers of ω distinct prime factors, such that this product involves no unity congruences nor does any prime factor p_j divide $p_i + 1$ or $p_i^2 + p_i + 1$, where p_i and p_j run over the prime factors of the order. The order for which there are 7^ρ abstract groups may be constructed as the product of the cubes of ρ distinct prime factors and ρ additional distinct primes, such that the factors may be written in 2ρ sets of the form $p_i^3 p_j$, and such that, for every set, p_j divides $p_i + 1$.[8] The order for which there are p groups, where p is of the form in the theorem, is shown to exist in table I. By choosing the prime factors in the orders for which there are 2^α, 3^β, 5^ω, 7^ρ and p groups so that theorem 2 is satisfied, the theorem follows.

COROLLARY. *There exists at least one order g for which there are precisely k finite abstract groups, for all integral values of k not exceeding* 120.

KANSAS STATE COLLEGE,
MANHATTAN, KANSAS.

ANNALS OF MATHEMATICS
Vol. 40, No. 1, January, 1939

THE REDUCTION OF POSITIVE QUATERNARY QUADRATIC FORMS

BY BURTON W. JONES

(Received April 13, 1938)

1. Introduction

The background of this problem is developed in an earlier paper.[1] We here complete the reduction referred to on page 518 as the "third problem." That is, we here determine conditions on the coefficients of a positive quaternary quadratic form which define a reduced form. This reduced form has the property that in every class of equivalent forms there is one and only one reduced form. In other words, every form is equivalent to one and only one reduced form.

These conditions are in the form of linear inequalities on the coefficients and fall into two groups. First, there are inequalities which are *completely symmetric*, that is, are independent of the numbering of the variables of the form. These conditions are, except for those mentioned in the previous paper, of a distinctly new type and possess many obvious advantages one of which is the brevity with which the conditions may be stated. Second, there are the usual inequalities which control permutations of the variables.

In the reduction, no distinction is made between *proper and improper equivalence*. That is, transformations of determinant -1 are used as freely as those of determinant $+1$. However if one desires a set of reduced forms of given determinant having the property that every form of that determinant is *properly* equivalent to one and only one form of the set, the set may be found as follows. First, find the set of reduced forms of the given determinant by the inequalities of this paper. Then for each reduced form which has no automorph of determinant -1 we will have another form obtained from it by, for instance, changing the signs of the coefficients of x_1x_2, x_1x_3 and x_1x_4 in the form. Hence the number of forms in what we might call the properly reduced set will be the number of forms of the reduced set plus the number of forms in that set without automorphs of determinant -1. In the table of automorphs at the close of the paper, the automorphs of determinant -1 are starred.

S. B. Townes, has, in an unpublished Chicago dissertation (1936), carried through the Eisenstein reduction for positive quaternary quadratic forms. Though it has the advantage of giving a leading coefficient which is the minimum of the form, etc., the proof is very much longer and the conditions on the form are more complex. The comparative shortness of this proof is due in large part to the symmetric conditions above referred to. Furthermore, the minimum of any reduced form of this paper can be obtained quickly.

[1] B. W. Jones, Annals of Mathematics, vol. 35 (1934), pp. 516–528.

We use the notations of sections 1 and 3 of the earlier paper and the results (and definitions) expressed in the theorems of sections 5 and 6. The diagrams there mentioned play a fundamental rôle. *Throughout this paper, i, j, p, m, n* are 1, 2, 3, 4, 5 in some order. In any paragraph different letters denote different numbers. By way of abbreviation we represent $-a_{ij}$ by ij, and $-a'_{ij}$ by ij'. The word "coefficient" refers to an a whose subscripts are not equal. Small unprimed Greek letters denote inequalities. If they are primed, double-primed, etc., it is understood that the equality may also be admitted. Thus, to say that $im\ \alpha'\ in$ and $pm + jm\ \alpha''\ pn + jn$ would mean that one of the four following conditions must hold: $im = in$, $pm + jm = pn + jn$, $im < in$ *and* $pm + jm < pn + jn$, $im > in$ *and* $pm + jm > pn + jn$. We would also write $im + pm + jm\ \alpha' + \alpha''\ in + pn + jn$ meaning that $im + pm + jm = in + pn + jn$ if and only if α' and α'' are both $=$ and that if they are not both $=$ the inequality has the same direction as α' or α''. $im\ \alpha'\ in$ is also written $in\ \underline{\alpha'}\ im$ and $(\underline{jip})$ means that the condition (jip) does not hold. To write $\alpha' = \beta'$ means that either they are both $=$ or they are both inequalities in the same direction. $\alpha' = \beta$ would mean that α' is an inequality (not an equality) in the same direction as β. If we are considering α', α'', β', β'' and write $\alpha = \beta$ we mean that one of the three following conditions hold: α' and α'' are $=$, β' and β'' are $=$, the inequality α' or α'' has the same direction as the inequality β' or β''. I_x, I_y, etc., denote transformations which are obtained from the identity by permuting its columns.

In order to establish the reduction we first make a form semi-reduced as in the earlier paper and then consider the transformations there given which take a semi-reduced form into a semi-reduced form. We consider first a permissible transformation A_{jip} and show that it reverses a certain condition, that is, the form satisfies the condition after the application of the transformation if it did not satisfy it before and vice versa. Hence the form is equivalent to one which satisfies the condition; but if it satisfies the condition the transformation will take it into a form which will not satisfy the condition unless it is an automorph or merely permutes the variables. We deal similarly with other permissible transformations.

The chief object of this paper is to prove that there is one and only one form in each class of (properly or improperly) equivalent forms satisfying the following conditions.

I. Symmetric Conditions

a) All coefficients $\leqq 0$

1. If $a_{ij} = a_{ip} = 0$ and if[2]

$(jip)_1$ $\quad |\, a_{im} \,|\, \alpha \,|\, a_{in} \,|$ then $|\, a_{jm} + a_{pm} \,|\, \alpha' \,|\, a_{jn} + a_{pn} \,|$.

[2] If $a_{im} = a_{in}$ we say that (jip) automatically holds; similarly below.

and if α' is $=$ either[3]

$(jip)_2$ $|a_{im}| < |a_{in}|$ and $|a_{jm}|\beta'|a_{jn}| \cdot\sim\cdot |a_{pn}|\beta'|a_{pm}|$[4] and

$|a_{jm}|\beta''|a_{pn}| \cdot\sim\cdot |a_{jn}|\beta''|a_{pm}|$ or $|a_{im}| > |a_{in}|$ and

$|a_{jn}|\beta'|a_{jm}| \cdot\sim\cdot |a_{pm}|\beta'|a_{pn}|$ and $|a_{jm}|\beta''|a_{pn}| \cdot\sim\cdot |a_{jn}|\beta''|a_{pm}|$.

2. If $a_{ij} = a_{ip} = a_{pm} = 0$ and all the other coefficients are negative and if

$(jipm)$
$$|a_{jn} + a_{jp}|\alpha''|a_{jm}|\alpha'|a_{jn} + a_{pn}| \text{ and}$$
$$|a_{im} + a_{mm}|\beta''|a_{jm}|\beta'|a_{in} + a_{mn}|,$$

where $\alpha'' + \alpha' = \beta'' + \beta'$ and the only pairs from α', β', α'', β'' which can be simultaneously equal are α' and β' or α'' and β''; if, furthermore, when α' or β' is $=$, α and β are $<$ and when α'' or β'' is $=$, α and β are $>$, then

$$|a_{in} + a_{mn}| < |a_{jm}| > |a_{jn} + a_{pn}|.$$

b) $a_{ij} > 0$ **and all other coefficients** $\leqq 0$

1. $a_{ij} + a_{ip} \leqq 0 \geqq a_{ij} + a_{jp}$ for every $p \neq i$ or j.
2. If $a_{ij} + a_{ip} = 0$

$[jip]_1$ $$|a_{jm} + a_{jn} + a_{ij}| \leqq |a_{pm} + a_{pn} + a_{ip}|.$$

$[jip]_2$ If the $=$ holds above and if $|a_{jm} + a_{ij}|\alpha|a_{pn}| \cdot\sim\cdot |a_{pm}|\alpha|a_{jn} + a_{ij}|$ then $|a_{pn}|\alpha'|a_{jn} + a_{ij}| \cdot\sim\cdot |a_{jm} + a_{ij}|\alpha'|a_{pm}|$.

$[jip]_3$ If the equality holds in the first line and α' is $=$ then $|a_{im}|\alpha''|a_{in}|$.

3. If $a_{ij} = -a_{ip} = -a_{im} (\neq -a_{in})$[5] then

$\{jin\}_1$ $$|a_{jp} + a_{jm} + a_{ij}| \leqq |a_{pn} + a_{mn} + a_{ip}|.$$

$\{jin\}_2$ If the $=$ holds and $|a_{jm} + a_{ij}|\sigma|a_{pn}| \cdot\sim\cdot |a_{mn}|\sigma|a_{jp} + a_{ij}|$, then

$|a_{pn}|\sigma'|a_{jp} + a_{ij}| \cdot\sim\cdot |a_{jm} + a_{ij}|\sigma'|a_{mn}| \cdot\sim\cdot |a_{jm} + a_{pn}|$

$\sigma'|a_{mn} + a_{jp}|$.

4. If $a_{pn} = a_{nm} = 0$ and all the conditions $[\cdots]$ (i.e. conditions 1 above) of the diagram hold as well as those obtained from them by interchanging a_{in} and a_{jn} and if $a_{in} \neq a_{jn}$, we say (pnm) is *applicable* and require

$(pnm)_1$ $|a_{jn}|\alpha|a_{in}|$ and $|a_{jm} + a_{jp}|\alpha'|a_{im} + a_{ip}|$

and if α' is $=$ then either

---

[3] Another statement of this condition is given near the beginning of section 2.

[4] That is $|a_{jm}|\beta'|a_{jn}|$ and $|a_{pn}|\beta'|a_{pm}|$ are, in virtue of α' being $=$, equivalent statements.

[5] 3. becomes identical with $[jin]$ if the equality holds.

$(pnm)_2$ $|a_{jn}| < |a_{in}|$ and a_{im} and a_{ip} lie between a_{jm} and a_{jp}

or $|a_{jn}| > |a_{in}|$ and a_{jm} and a_{jp} lie between a_{im} and a_{ip}.

(If α' is $=$ and $a_{ip} = a_{jm}$ or a_{jp}, the condition automatically holds. See also A 11 in section 3.)

5. If $a_{ij} + a_{ip} = 0 = a_{pn}$ and if in the complete diagram every $\{\cdots\}$ (condition 3. above) and every $[\cdots]$ hold and if *none* of the following hold:

1) $|a_{mn} + a_{jn}| \geqq |a_{in}|$,

2) $a_{jn} = a_{ip}$ and $|a_{jp}| > |a_{jm}|$,

3) $a_{im} = a_{ip}$ and $|a_{jp} + a_{pm}| \geqq |a_{in}|$,

4) $a_{jp} = a_{ip}$ and $|a_{im} + a_{mn}| \geqq |a_{jm}|$,

5) $a_{jm} = a_{ip}$ and $a_{mn} = 0$

then we say $[npij]$ *is applicable* or *applies* and require

$[npij]_1$ $|a_{pm} + a_{mn} + a_{ip}| \geqq |a_{in} + a_{jm} + a_{ij}|$,

$[npij]_2$ if the $=$ holds, then $|a_{pm}| \leqq |a_{in} + a_{ij}| \cdot\sim\cdot |a_{jm}| \leqq |a_{mn} + a_{ip}|$,

$[npij]_3$ if both equalities hold, then $|a_{jp}| \leqq |a_{im}|$.

(See A 9 in section 3.)

6. If $a_{ij} + a_{ip} = a_{ij} + a_{jn} = a_{pn} = 0$, if all conditions $\{\cdots\}$ and $[\cdots]$ hold and if

1) $|a_{jm}| < |a_{jp}|$ and $|a_{im}| < |a_{in}|$, and

2) $a_{im} \neq a_{ip}$ *or* $a_{im} = a_{ip}$ and $|a_{pm} + a_{jm}| < |a_{in}|$, we say that $\{npij\}$ or, what is the same thing, $\{pnji\}$ *applies* and require

$\{npij\}_1$ $|a_{jp} + a_{in} + a_{ij}| \leqq |a_{pm} + a_{nm} + a_{ip}|$

$\{npij\}_2$ if the $=$ holds and $|a_{jp} + a_{ij}|\alpha|a_{pm}| \cdot\sim\cdot |a_{mn} + a_{ip}|\alpha|a_{in}|$ then $|a_{jp} + a_{ij}|\alpha'|a_{mn}| \cdot\sim\cdot |a_{pm} + a_{ip}|\alpha'|a_{in}|$,

$\{npij\}_3$ if α' is $=$, then $|a_{im}|\alpha''|a_{jm}|$.

(See A 10 in section 3.)

NOTE: In any given diagram not more than one of the following three conditions can be simultaneously applicable $[npij]$, $[pnji]$, $\{npij\}$.

c) $a_{ij} > 0$, $a_{nm} > 0$ **and all other coefficients** < 0

1. $a_{ij} + a_{ip} \leqq 0 \geqq a_{ij} + a_{jp}$ for $p \neq i$ or j.
   $a_{mn} + a_{np} \leqq 0 \geqq a_{mn} + a_{mp}$ for $p \neq m$ or n.
2. $a_{ij} + a_{mn} + a_{ps} \leqq 0$ for $p = i$ or j and $s = m$ or n.
3. If $a_{nm} + a_{jn} + a_{ij} = 0$, then

$(ijnm]_1$ $\quad | a_{jp} + a_{np} + a_{jn} | \geqq | a_{jm} + a_{in} |$

and if the $=$ holds

$(ijnm]_2$ $\quad | a_{jp} | \alpha' | a_{in} + a_{nm} | \cdot\sim\cdot | a_{jm} + a_{ij} | \alpha' | a_{np} |$ and
$| a_{np} | \alpha'' | a_{in} + a_{ij} | \cdot\sim\cdot | a_{jm} + a_{nm} | \alpha'' | a_{jp} |$.

$(ijnm]_3$ $\quad$ If α' is $=$ and α'' not then $| a_{ip} + a_{ij} | \alpha''' | a_{mp} + a_{mn} |$;
if α'' is $=$ and α' not then $a_{mn}\alpha^{IV}a_{ij}$;
if α' and α'' are $=$ then α''' and α^{IV} conditions hold.

4. If $a_{ij} + a_{ip} = 0$ and *none* of the following hold
   1) $a_{ij} + a_{jn} + a_{nm} = a_{ij} + a_{im} + a_{mn} = 0$,
   2) $a_{ij} + a_{jm} + a_{nm} = a_{ij} + a_{in} + a_{mn} = 0$,
   3) the diagram contains J_5' (see section 4)

   then we impose condition *b*) 2.

II. Unsymmetric Conditions

1. $a_{ii} \leqq a_{jj}$ if $i < j$.
2. If $a_{ii} = a_{jj}$ where $i < j$ and if $| a_{ip} | \alpha' | a_{jp} |$, $| a_{im} | \beta' | a_{jm} |$, $| a_{in} | \gamma' | a_{jn} |$ where α', β' and γ' are not all equalities, then
   1) if none of α', β', γ' are $=$, two or more are $<$,
   2) if γ' is $=$ but α' and β' not, then α' is $<$ if $p < m$ while β' is $<$ if $m < p$; and a similar condition if β' is $=$ or α' is $=$,
   3) if two of α', β', γ' are $=$, the third is $<$.

We shall find useful the following relationships among the fundamental transformations, which can easily be verified:

$$A_{jip}A_{pnm} = B_{ipnm}\mathrm{I}_{in,jm} \; ; \; A_{jip}A_{ijp} = A_{ipj}I_{mn} \; ; \; A_{jip}A_{ipm} = B_{jipm} \; ;$$
$$A_{jim}A_{jin}A_{jip} = I_{-i,n} \; ; \; \mathrm{I}_{-p,j}A_{jip}A_{mpn}\mathrm{I}_{mn} = C_{jip,nm} \; ; \; \mathrm{I}_{-i,m}A_{nip}A_{jim} = E_{ij,pnm} \; ;$$
$$I_{-jp,m}B_{ipnj} = A_{pij}B_{ipnj}I_{in} \; ; \; I_{-ij,m}B_{pinj}I_{-p,n}A_{mpi} = I_{-p,n}A_{mpi}I_{pn,jm} \; ;$$
$$A_{ijn}B_{ipmn}I_{ij,pn} = D_{ijnmp} \; ; \; B_{ijmn}A_{mnp} = A_{mnp}\mathrm{I}_{ip,jn} \; .$$

2. The symmetric conditions when $a_{ij} \leqq 0$ for all i and j, $i \neq j$

If $a_{ij} = a_{ip} = 0$, that is $ij = ip = 0$, the earlier paper (p. 523) shows that A_{jip} and the transformations obtained from it by permuting the columns take a semi-reduced form into a semi-reduced form. We shall use especially

A_{jip} which interchanges[6] im and in,
$A_{jip}I_{mn}$ which interchanges jm and jn, pn and pm,
$A_{jip}I_{jp,mn}$ which interchanges jm and pn, jn and pm.

These transformations leave unaltered all coefficients not mentioned. In virtue of the first line, A_{jip} is an automorph if $im = in$ but otherwise changes the first inequality of $(jip)_1$ in the introduction and leaves the second. Hence, unless α' is $=$, if the coefficients of f satisfy $(jip)_1$, the application of A_{jip} will

[6] This may be seen by applying A_{312} to φ and finding φ'.

give a form f' whose coefficients do not satisfy $(jip)_1$ and if the coefficients of f do not satisfy $(jip)_1$ the coefficients of f' will. If α' is $=$, suppose, for instance, $jm\ \beta'\ jn$ and $jm\ \beta''\ pn$; then if $im < in$ the application of A_{jip} will reverse the inequality but leave the rest of the line and hence deny the condition unless β' or β'' is $=$, but if $im > in$, the application of A_{jip} will yield a form which does satisfy the condition. If, on the other hand, $jn\ \beta'\ jm$ and $jm\ \beta''\ pn$ the same procedure may be carried through. If in $(jip)_2$ β' is $=$, $A_{jip}I_{mn}$ is an automorph while if β'' is $=$, $A_{jip}I_{jp,mn}$ is an automorph. Hence, unless one of the three transformations is an automorph, the application of A_{jip} to a form which satisfies (jip) yields a form which denies it and vice versa. In other words, every form for which $ji = ip = 0$ is equivalent to a form which satisfies (jip) and A_{jip} will not take it into another form satisfying (jip) unless one of the three transformations is an automorph.

Furthermore, inspection shows that the condition (jip) remains unaltered if j and p are interchanged, or if m and n are interchanged. These interchanges together with their combination are the only permutations of subscripts which leave $ji = ip = 0$ unaltered. Any other permutation would merely replace $ijpmn$ by $i'j'p'm'n'$ and hence would still leave the condition (jip), though now it would be in terms of j', i' and p'. As for the case when, for example, $A_{jip}I_{mn}$ is an automorph, it may be seen that then the effect of A_{jip} on the form is the same as that of I_{mn} and therefore A_{jip} will be taken care of when we later provide for equivalence under the transformation I_{mn}. Also note that $A_{jip}I_{mn}$ and $A_{jip}I_{jp,mn}$ will have the same effect on (jip) as will A_{jip}.

The further development is simplified by noting that $(jip)_2$ may be written:

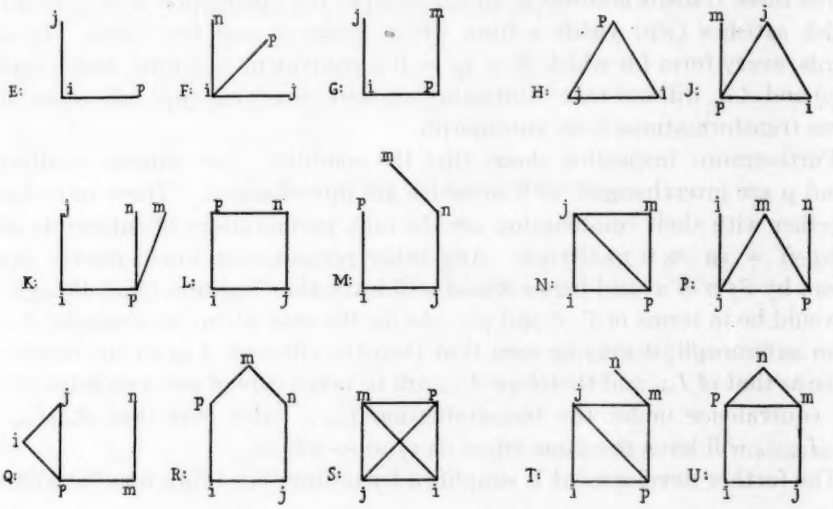
if $im < in$ then pn and jn are between jm and pm, while
if $im > in$ then jm and pm are between jn and pn

and if in either line one of the second two is equal to one of the last two, the condition automatically holds since $A_{jip}I_{mn}$ or $A_{jip}I_{jp,mn}$ is an automorph.

Though the above shows that (jip) can be satisfied in a single case it is conceivable that in a form in which several such conditions applied we might deny the condition for one pair of vanishing coefficients while in the process of satisfying the conditions for another pair. It therefore becomes necessary to investigate all the varieties of combinations of such conditions which may occur and to show finally that every form is equivalent to a form satisfying a condition (jip) for *every* vanishing pair of coefficients having a subscript in common. We will then know from the above theory that, once a form satisfies all such conditions, the application of any permissible A_{jip} will alter one condition if it alters the form unless its application has the same effect on the form as a permutation of variables. After this is done we proceed to dispose of the transformations B, C and D.

In this section, by the diagram of a form we mean a polygon (not necessarily closed) containing a line ji for each coefficient ji which vanishes. That is, if $ji = ip = mp = mj = 0$ we would have a "rectangle" with vertices i, j, p and m.

(Obviously the shape of the polygon has no significance.) In finding all the possible diagrams it should be noted that the corollaries 1 and 2 of the earlier paper imply that not more than six lines can occur nor can one point be the end of more than three lines if the determinant of f is to be different from zero. If there are three lines, for example, the distribution of lines among the vertices is one of the following: (3, 1, 1, 1), (2, 2, 1, 1) or (2, 2, 2). Proceeding similarly for the other cases we have the diagrams of set I together with L_1, which is L with the addition of line pm.



Consider the diagram F. In the discussion of this diagram and those below, though we assume $im > 0$ (which follows from the form being positive) we make no assumption about the appearance of other lines in the diagram. We show that, whatever the diagram containing F is, we can satisfy the conditions associated with the F portion of the diagram. We list the following conditions:

(jip) $\quad im > in$ and $jm + pm\,\alpha'\,jn + pn,$

(pin) $\quad im > ij$ and $pm + nm\,\beta'\,jp + jn,$

(jin) $\quad im > ip$ and $jm + nm\,\gamma'\,jp + np.$

Since $A_{jip}I_{mn}$ does not alter the diagram we may assume from previous discussion that (jip) holds. Suppose (pin) does not. Then α' is $\geqq \beta'$ is $\leqq$. Let $np + pm\,\delta'\,jn + jm$ and if $A_{pin}I_{jm}$ is applied the condition (pin) will hold in the resulting form. The transformation leaves unaltered the first part of every condition. We have for the new form in terms of the old coefficients:

$(jip)'$ $\quad jm + jp\,\alpha' + \underline{\beta}'\,mn + pn$

$(jin)'$ $\quad jm + jn\,\underline{\delta}'\,pm + pn.$

If α' and β' are $=$, $(jip)_2$ is $jn\ \epsilon'\ jm\ \epsilon''\ pn$, $jn\ \epsilon''\ pm\ \epsilon'\ pn$; $(\underline{pin})_2$ is $nm\ \lambda\ jp\ \lambda\ pm$, $nm\ \lambda\ jn\ \lambda\ pm$ which implies $\lambda = \epsilon''$ and hence $(jip)_2'$ is $mn\ \lambda + \epsilon'\ jm\ \epsilon''\ pn$. Hence $(jip)'$ holds. If (jin) holds, we see from symmetry that $(jin)'$ would hold. In any case we have shown that after the application of $A_{pin}I_{jm}$ two of the conditions associated with the diagram hold. Hence we may now begin with (jip) and (jin) holding. Then, if (pin) does not hold, the above discussion may be applied again to complete the proof.

Diagram G. If $jp = pn$ and (jip) does not hold, apply A_{jip} which alters neither the diagram, jp nor pn and hence (jip) and (ipm) for the new form will hold. Thus it remains to consider the case when $im \neq in$ and $jp \neq pn$. Apply A_{ipm} if necessary to make (ipm) hold. We must then consider the case when (jip) does not hold. In this case we have $im\ \alpha\ in$ and $jm\ \underline{\alpha}'\ pn + jn$, $jp\ \beta\ pn$ and $jm\ \beta'\ in + mn$. First, if $\alpha = \underline{\beta}$ use A_{jip} to fix (jip). This leaves the third inequality and changes the fourth to $jm\ \gamma'\ im + mn$. But $jm\ \beta'\ in + mn\ \underline{\alpha}\ im + mn$. Hence $\gamma' = \beta' + \underline{\alpha}$ and (ipm) is also satisfied by the new form. Second, if $\alpha = \beta$ the application of A_{jip} yields a new form in which $im'\ \underline{\beta}\ in'$, $jp'\ \beta\ pn'$ and $(jip)'$ holds. Then if $(ipm)'$ does not hold we may interchange the rôles of (jip) and (ipm) and proceed as in the first case. Note that here, as in F, we did not assume that no other lines were zero in proving that (jip) and (ipm) could be satisfied simultaneously.

Diagram H. If two pairs of $im, in; jm, jn; pm, pn$ were $=$ we could, as in G, use the proper transformation if necessary to fix the condition associated with the third and the conditions associated with the others would automatically hold. Hence we suppose that (jip) holds and (ipj) does not and that $im\ \alpha\ in$ while $pm\ \beta\ pn$. It may be shown by methods similar to those above that if $\beta = \underline{\alpha}$, A_{ipj} makes all conditions hold while if $\beta = \alpha$, $A_{jip}A_{ipj} = A_{pji}I_{mn}$ may be used.

Diagrams J and K. In fixing the conditions associated with the lines of J or K terminating in p we may use a transformation A_{ipm} which interchanges the diagrams J and K. A_{jpm} would have the same effect except that the letters would be changed. Hence we assume from the theory for F that the conditions (jpm), (ipm), (jpi) hold for J and (ipn), (ipm), (npm) for K. Denote the second inequalities in the conditions for J by α', β' and γ' respectively where each is $\leqq$. First, suppose that (pji) does not hold in J, that is $jm\ \delta\ jn$, $im\ \underline{\delta}'\ pn + in$. The inequalities of (pij) are $im\ \epsilon'\ in$ and $jm\ \lambda'\ pn + jn$. Write $jn + im\ \mu'\ jm + in$. If μ' is $\leqq$ we have $\epsilon' = \mu' + \gamma'$. Hence ϵ' is $\leqq$ and since $pn \neq 0$, $\underline{\delta}'$ is $<$ and therefore δ' is $>$. Apply A_{pji} and see that all the conditions hold for the new form with no equalities holding. On the other hand, if μ' is $>$, $\delta = \gamma' + \underline{\mu}'$ shows that δ is $<$ and hence ϵ' is $>$ while λ' is $<$. Hence (pij) is denied. The interchange of j and i reverses μ' and hence we may use the immediately preceeding case with i and j interchanged to complete the theory. Second, consider K where the second inequalities of the conditions (ipn), (ipm), (npm) are denoted by ρ', $\underline{\beta}'$ and σ' each being $\leqq$. From symmetry assume $im > in$ and if (jip) is denied $jm\ \tau'\ jn$ where τ' is $\leqq$. If A_{jip} fixes (ipm) all the other conditions

will hold and the diagram is left unaltered. On the other hand, if it does not $B_{jipm} = A_{jip}A_{ipm}$ will fix (ipm), will transform the diagram to J in which, in terms of the old coefficients, the new conditions (ipj) and (jpm) are $pm < jp$ and $in + jm < im + jn$; $jp > pi$ and $jn + mn > ni$, the last resulting from $\rho' \leqq$. We then proceed as above.

Diagram L. The transformations used to fix the conditions (pij) and (ijn) do not alter the diagram. Hence we may, from G, suppose that the conditions (pij) and (ijn) hold. It may be shown that application of A_{jnp} and then, if necessary, A_{npi} fixes all the conditions.

Diagram M. We may assume from G, as above, that (pij) and (ijn) hold. It may be shown that application of A_{jnm} fixes all conditions.

Diagrams N, P and L_1. The transformations adjusting the conditions associated with the vertex p of the diagrams permute the diagrams. But condition (ipn) for L_1 cannot hold and thus we need to consider only N and P. Hence we assume that in $N(jpm)$ holds which implies $im\ \alpha'\ jn + mn$ where α' is $\leqq$ and (jpi) holds which implies $im\ \beta'\ jn + in$ where β' is $\leqq$. (ipm) and (ijm) hold automatically. (jip) holds if and only if $im\ \gamma'\ in$ where γ' is $\leqq$. (jmp) holds if and only if $im\ \epsilon'\ mn$ where ϵ' is $\leqq$. First, if ϵ' is $>$, A_{jmp} will fix (jmp) and (jpm), (ipm) and (jpi) still hold. (jip) will hold after the transformation if it did previously. In the resulting form $im' \leqq mn'$. If then $(jip)'$ does not hold apply A_{jip} to f' to yield a form in which all the conditions hold. If (jip) and (jmp) hold, so do (pji) and (pjm). In P we assume as above that (jpn) and (jpm) hold. Then A_{mjp} takes P into N with i and n interchanged. It may be seen that $(jpn)'$ and $(jpm)'$ hold and the theory for N carries through as above.

Diagram Q. Assume as above that (jpm), (ipm), (jpi) hold, since all the transformations involved leave the form of the diagram unaltered. If $in < im$ use A_{jip} to give f' in which $in' \geqq im'$ and the conditions above hold. If in f', $jn' < jm'$ use A_{pji} which yields f'' in which $in'' \geqq im''$ and $jn'' \geqq jm''$ and all the conditions except perhaps (nmp) hold. If (nmp) does not hold use A_{nmp}.

Diagram R. Assume (pij), (ijn) and (jnm) hold and write their first inequalities in order $im\ \alpha'\ in$, $jp\ \beta'\ jm$, $np\ \gamma'\ ni$ and denote their second inequalities by α'', β'', γ'' respectively. Since an interchange of p and m, i and n interchanges the first and third conditions and reverses both inequalities of the second we assume $\alpha = \gamma = \beta$ or $\alpha = \underline{\gamma}$. In the first case conditions (nmp) and (mpi) hold since $jp\ \beta'\ jm\ \alpha''\ pn\ \beta''\ im\ \alpha'\ in$. In the second case it may be seen that $\gamma = \beta$ and A_{nmp} will fix (nmp) if it is denied and all the other conditions except perhaps $(mpi)'$ which holds if and only if (mpi) does. If $(mpi)'$ does not hold apply the above theory with p and m, i and n interchanged to complete the proof.

In S all the conditions automatically hold. A_{jmp} takes T into S and $B_{ipmj} = A_{ipm}A_{pmj}$ takes U into S.

We now wish to find under what conditions B_{jipm} leaves the conditions (jip)

and (ipm) in a diagram containing G. Note that $A_{jip}A_{ipm} = B_{jipm}$ and hence this transformation interchanges im and in as well as pn and jp. Suppose

(jip) $\quad im\ \alpha\ in$ and $jm\ \alpha'\ jn + pn$

(ipm) $\quad jp\ \beta\ pn$ and $jm\ \beta'\ in + mn$

$(jip)'$ $\quad in\ \underline{\alpha}\ im$ and $jm\ \underline{\alpha}''\ jn + jp$

$(ipm)'$ $\quad pn\ \underline{\beta}\ jp$ and $jm\ \underline{\beta}''\ im + mn.$

Note that $\beta' + \beta'' = \alpha$, $\alpha' + \alpha'' = \beta$. When $A_{jip}I_x$ or $A_{ipm}I_y$ is an automorph we need not consider B_{jipm} since, for example, $A_{jip}I_{mn}A_{ipn}I_{mn} = B_{jipm}$ and if $A_{jip}I_{mn}$ is an automorph the application of B_{jipm} has the same effect on the form as $A_{ipn}I_{mn}$ which has already been provided for. Hence we suppose $im \neq in$ and $jp \neq pn$ and if α' is $=$, no equality in $(jip)_2$ need be considered and hence since pm cannot lie between pn and jn we have:

if α' is $=$ then α is $<$ and $jm > pn, jn > pm = 0,$

if β' is $=$ then β is $<$ and $jm > in,\ mn > ij = 0,$

if α'' is $=$ then $\underline{\alpha}$ is $<$ and $jm > jp,\ jn > pm = 0,$

if β'' is $=$ then $\underline{\beta}$ is $<$ and $jm > im,\ mn > ij = 0.$

Note that in the first line $pn \neq 0 \neq jn$ and $\alpha <$ imply $jm > pn$ and jn. Also α' and β'' are not both $=$, β' and α'' are not both $=$. The condition $in + nm\ \underline{\beta'}$ $jm\ \alpha'\ jn + pn$ becomes after the transformation B_{jipm}, $im + nm\ \beta''\ jm\ \underline{\alpha''}\ jn + jp$. Note that if α' or β' is $=$, α'' and β'' are $<$ while if α'' or β'' is $=$, α' and β' are $>$. Hence either before or after the application of the transformation but not both it will be true that $in + nm < jm > jn + pn$ and we impose the condition of the introduction.

We proceed to show that the above condition need be imposed only if all the coefficients not occurring in G are different from zero. We show that otherwise if all the $(\cdots)$ conditions hold and B_{jipm} applies, it must take it into a form in which not all the $(\cdots)$ hold. First $jm \neq 0$. Second, if $pn = 0$, the condition implies that β is $>$ and thus α is $>$ and neither α' nor β' is $=$. B_{jipm} takes K into J in which if $(ijp)'$ held, $jn < jm$ and $im + jp \leqq in$ which contradicts $im > in$. Third, if $jp = 0$, then $jp < pn$ and the condition implies $jn \leqq jm \leqq jn + pn$ and $im < in$ while if (pji) holds $jm \geqq jn$ and $im \geqq pn + in$ which is a contradiction. Fourth, if $mn = 0$, the condition implies $im\ \beta''\ jm\ \beta'\ in$ and neither β' nor β'' is $=$ if the form is positive. If (pmn) held we would have $mi\ \beta\ mj$ and $ni\ \gamma'\ pj + .ij$. But $pj + nj\ \alpha''\ jm\ \beta'\ in$ and hence $\gamma' = \underline{\beta'} + \underline{\alpha''}$ and (pmn) does not hold. Finally, note from symmetry that we have considered all the typical cases. This disposes of condition $(jipm)$.

We now prove that if $C_{jip,nm}I_{ji}$ is applied to a form whose diagram contains K and whose coefficients satisfy the conditions already imposed, the coefficients

of the resulting form do not satisfy all the conditions imposed unless the transformation has the same effect as a previous transformation. It leaves the coefficients of the form except for the following changes: $jn' = in$, $in' = jm$, $jm' = im$, $im' = jn$. Hence (pij) is $im\ \alpha'\ in$ and $jm\ \alpha''\ jn$ while $(pij)'$ is $jn\ \underline{\alpha''}\ jm$ and $im\ \alpha'\ in$ which is false unless α'' or α' is $=$. First if α' is $=$ note that $C_{jip,nm}I_{ij} = I_{-p,j}A_{jip}A_{mpn}I_{mn,ji}$ and since $I_{-p,j}$ and A_{jip} are automorphs, the effect of $C_{jip,nm}I_{ji}$ is the same as that of $A_{mpn}I_{mn,ij}$ which by our previous theory is an automorph or alters some condition imposed. Second, if α'' is $=$ write $C_{jip,mn}I_{ij} = I_{-p,j}A_{jip}I_{mn}I_{mn}A_{mpn}I_{mn,ij} = I_{-p,j}A_{jip}I_{mn}A_{mpn}I_{ij}$ and since $A_{jip}I_{mn}$ and $I_{-p,j}$ are automorphs we proceed as above.

Finally we prove that if D_{ijnmp} is applied to a form whose diagram contains R and whose coefficients satisfy the conditions already imposed, the coefficients of the resulting form do not satisfy all the conditions imposed unless the transformation is an automorph multiplied by a transformation previously considered. The transformation leaves the coefficients of the form except for the following changes $pn' = jm$, $jm' = jp$, $jp' = in$, $in' = pn$. Hence since (pij) is $in\ \alpha'\ im$ and $pn\ \alpha''\ jm$ while (ijn) is $jm\ \beta'\ jp$ and $im\ \beta''\ np$ we see that $(pij)'$ is $pn\ \underline{\beta''}\ im$ and $jm\ \beta'\ jp$ which does not hold unless β' or β'' is $=$. First if β' is $=$, A_{ijn} is an automorph and we use $D_{ijnmp} = A_{ijn}B_{ipmn}I_{ij,pm}$. Second, if β'' is $=$, $A_{ijn}I_{in,pm}$ is an automorph and we use $D_{ijnmp} = A_{ijn}I_{in,pm}B_{ipmn}I_{inj}$.

3. Symmetric conditions for $a_{ij} > 0$ and all other coefficients $\leqq 0$

(Throughout this section i and j are understood to be the subscripts in some order of the positive coefficient). Condition b) 1. of the introduction arises from the earlier paper.

If $ji + ip = 0$ we shall have occasion to use two transformations:

$A_{jip}I_{jp}$ which interchanges: im and in; pm and $jm + ij$, that is, $pm + ip$ and jm; pn and $jn + ij$, that is, $pn + ip$ and jn.

$A_{jip}I_{jp,mn}$ which interchanges: pn and $jm + ij$, that is, $pn + ip$ and jm; pm and $jn + ij$, that is, $pm + ip$ and jn.

To establish the condition $[jip]$ of the introduction it can be seen that $A_{jip}I_{jp,mn}$ reverses $[jip]_1$, reverses the first but not the second of the inequalities of $[jip]_2$ and leaves the inequality of $[jip]_3$. If $jm + ij = pn$ and $pm = jn + ij$ the transformation is an automorph. If α' and α'' are $=$, the transformation $A_{jip}I_{jp}$ is an automorph. Hence, as in section 2, any form for which $ij + ip = 0$ can be transformed into a form which satisfies $[jip]$ and if one of the above transformations is applied to a form which satisfies $[jip]$, either one is an automorph or the transformation yields a form which does not satisfy $[jip]$. It should be noted that $[jip]$ is left unaltered by an interchange of m and n for the effect of such an interchange is to leave $[jip]_1$ and reverse *all* the other inequalities. Also, we could have used $A_{jip}I_{jp}$ above in place of $A_{jip}I_{jp,mn}$. Note that the condition $[jip]$ is different from $[ijp]$.

As in section 2 it becomes necessary to investigate all possible diagrams and to consider the other permissible transformations listed in the first paper. This time the diagrams consist of three types of lines: ij which represents the

positive coefficient; some, none or all of ip, im, in, jp, jm, jn which represent coefficients which when added to ij give zero; and some, none or all of mn, pm, nm which represent zero coefficients. We call the last three, *zero lines*. Note that, for instance, $A_{jip}I_{jp,mn}$ interchanges the lines jn and pm, jm and pn of any diagram. We can reduce the number of diagrams to be considered by means of Lemma 1 in which we use the following two transformations:

$I_{-i,n}A_{mip}I_{jn}$ which is permissible from the earlier paper if $ip = im = -ij$ and interchanges jm and $mn + im$, that is, $jm + ij$ and mn as well as jp and $ip + pn$, that is, pn and $jp + ij$. It interchanges the lines mn and jm, jp and pn of a diagram.

$B_{jipn}I_{ip,jn}$ which is permissible from the earlier paper if $ip + ij = 0 = pn$ and interchanges jp and im; pm and $in + ij$, that is, $pm + ip$ and in; nm and $jm + ij$, that is, $nm + ip$ and jm.

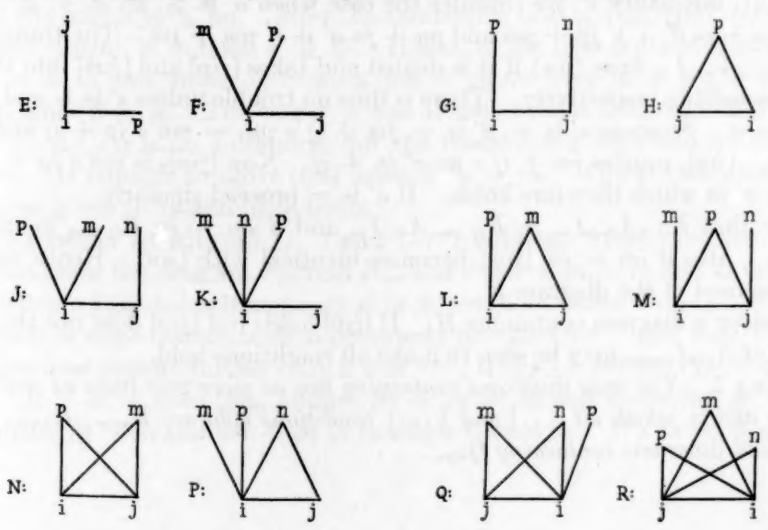
LEMMA 1. *Every diagram containing two or more zero lines may be taken by a permissible transformation into a diagram containing fewer zero lines with the exceptions below.*

First *if* $ji + ip = pn = pm = 0$, $A_{jip}I_{jp,mn}$ *reduces the number of zero lines unless* $jn = jm = ip$. *In this case* $I_{-j,p}A_{mjn}I_{ip}$ *is effective unless* $im = in = ip$. $jn = jm = im = in = ip$ *is the first exception.*

Second, *if* $ji + ip = pn = mn = 0 \neq pm$. *Use* $A_{jip}I_{jp,mn}$ *unless* jn *or* $jm = ip$. *If* $nj = ip$ *use the above with* p *and* n, i *and* j *interchanged. If* $in = ip$ *we may proceed similarly. If* $jm = ip = im$, $I_{-i,n}A_{mip}I_{jn}$ *will decrease the number of zero lines unless* $jp = ip$. *Hence the only exceptions are the first exception and* $jn \neq jm = ip \neq in$ *where none or both of* im *and* jp *are equal to* ip.

This lemma, though it does eliminate from consideration many diagrams does not insure us against being forced to consider a diagram such as E_{pnm} below which necessarily arises in the course of the discussion.

In listing the diagrams of set II we picture only those without zero lines and indicate any zero lines by subscripts. For instance if in G, $pm = 0$ we designate the diagram by G_{pm}. If $pm = 0 = mn$ we call it G_{pmn}.



Consider the diagram F. Suppose $[jim]$ holds but $[jip]$ does not. Then

$$jp + jn + ij\ \alpha'\ pm + mn + ip \qquad \text{where } \alpha' \leqq,$$

$$pm + pn + ip\ \beta'\ jm + jn + ij \qquad \text{where } \beta' \leqq.$$

Note that $jp + pn\ \alpha' + \beta'\ jm + mn$. Application of $A_{jip}I_{jp,mn}$ fixes $[jip]$ and gives $[jim]_1'$: $jp + pm\ \gamma'\ jn + mn$ which condition holds if γ' is $<$. First, if γ' were $=$ $[jim]'$ would be satisfied unless $jp + ij\ \delta\ mn$ and $pm\ \delta'\ mn$ where if δ is $=$, $in'\ \delta\ ip'$, that is, δ is $>$. Consider this case below. Second, if γ' were $\geqq$ and, when the $=$ holds, the condition in the previous sentence holds, apply $I_{-i,n}A_{mip}I_{jn}$ to the original form which takes $[jim]$ into $pn + jn\ \alpha' + \beta' + \underline{\gamma'}$ $pm + jm$ which is thus satisfied unless perhaps α', β', γ' are all $=$. In that case $jp + ij\ \delta\ mn$ and $[jim]_2$ implies $nm\ \delta''\ jn + ij$ where if δ'' is $=$, $ip\ \delta'''\ in$. $[jip]_2$ denied implies $jm + ij\ \epsilon\ pn$, $jn + ij\ \epsilon'\ pn$ and if ϵ' is $=$, $in\ \epsilon\ im$, that is, ϵ is $>$. $\alpha' + \beta' =$ implies $\epsilon = \delta$. Also $pm\ \delta' + \delta'' + \epsilon'\ pn$ and therefore $jn\ \delta' + \delta'' + \epsilon'\ jm$. Now if $[jim]_2'$ is denied $pn\ \tau\ jm + ij$, $jn\ \tau'\ jm$ and if τ' is $=$, $in > ip$ and τ is $>$. Then $\tau = \underline{\epsilon}$ which contradicts $jn\ \delta' + \delta'' + \epsilon'\ jm$ unless τ' is $=$ which would imply that τ is $=$. Also the transformation takes $[jip]$ into $mn + jn\ \underline{\gamma'}\ pm + jp$ and $[jip]'$ holds unless perhaps γ' is $=$. In that case $mn\ \underline{\delta}\ jp + ij$, $jp\ \underline{\delta'}\ jn$ and if δ' is $=$, $im < in$ and $\underline{\delta}$ is $<$. Hence $[jip]'$ holds.

Hence we have shown that in diagram F or any diagram containing it $[jip]$ and $[jim]$ may be satisfied simultaneously. For this diagram $I_{-i,n}A_{mip}I_{jn}$ is also permissible. It may be seen to reverse the inequality of $\{jin\}_1$ and the second but not the first of $\{jin\}_2$. If the equality in $\{jin\}_1$ holds and $jm + ij = pn$, $I_{-i,n}A_{mip}I_{jn,mp}$ is an automorph while if the equality in $\{jin\}_1$ holds and σ' is $=$, $I_{-i,n}A_{mip}I_{jn}$ is an automorph. If p and m are interchanged $\{jin\}_1$ remains and *all* the inequalities of $\{jin\}_2$ are reversed. Hence this condition is connected with this transformation in the same way that $[jip]$ is with $A_{jip}I_{jp}$.

We seek now to show that in the diagram in which $[jip]$ and $[jim]$ hold, $\{jin\}$ may also be made to hold without disturbing the first two conditions. Calling the $\{jin\}_1$ inequality κ', we consider the case when α' is $\leqq$, $\beta' \geqq$, $\kappa' \geqq$. Note that $mn + jn\ \underline{\beta'} + \underline{\kappa'}\ jp + pm$ and $pn + jn\ \alpha' + \underline{\kappa'}\ pm + jm$. The transformation $I_{-i,n}A_{mip}I_{jn}$ fixes $\{jin\}$ if it is denied and takes $[jip]$ and $[jim]$ into the last two inequalities respectively. There is thus no trouble unless κ' is $=$ *and* one of β', α' is $=$. Suppose κ' is $=$, β' is $=$, $jm + ij\ \sigma\ pn \cdot\sim\cdot mn\ \sigma\ jp + ij$ and $jp + ij\ \sigma\ pn$. $[jip]_2$ implies $jm + ij\ \sigma\ pn\ \sigma'\ jn + ij$. Now $[jip]_2'$ is $mn\ \sigma\ jp + ij$ and $jp\ \sigma + \sigma'\ jn$ which therefore holds. If α' is $=$ proceed similarly.

Note that $I_{-i,n}A_{mip}I_{jn} = E_{ij,pmn}A_{nij}I_{jn}$ and if $im = in$ $E_{ij,pmn}$ is an automorph. Also if $im = in$ $\{jin\}$ becomes identical with $[jin]$. Hence we have also disposed of the diagram K.

Consider a diagram containing H. If $[ijp]$ holds but $[jip]$ does not the application of $A_{jip}I_{jp,mn}$ may be seen to make all conditions hold.

LEMMA 2. *The only diagrams containing two or more zero lines as well as the line ip and in which all* $[\cdots]$ *and* $\{\cdots\}$ *conditions hold are* E_{pnm}, H_{pnm}, G_{pmn}, N_{pnm} *and diagrams containing* Q_{mpn}.

PROOF: If $pn = pm = 0$, $[jip]$ implies $jm = jn = ip$ and then $\{ijp\}$ implies $in = im = ip$ and the diagram contains Q_{mpn}. Hence it remains to consider only diagrams in which $pn = nm = 0$ and $in \neq ip \neq jn$. In F_{pnm} and L_{pnm} $\{jin\}$ implies $jp = jm = ip$ and hence the diagrams are or contain N_{pnm}. The other diagrams in which $pn = nm = 0$ and $in \neq ip \neq jn$ are listed in the lemma.

In diagrams in which $pn = nm = 0$ we use the condition (pnm) which is the same as the condition (jip) (with n and i, j and m interchanged) of section 2 and which has there been established. However we need impose it only when a form satisfies all the conditions $[\cdots]$ and $\{\cdots\}$ and when application of A_{pnm} leaves the conditions still holding. It may be seen that in any diagram containing N_{pnm} or Q_{mpn}, the condition (pnm) or (mpn) automatically holds. *Hence the condition* $(\cdots)$ *is to be considered only in the case of the first three diagrams listed in Lemma* 2.

We consider next the conditions associated with $B_{jipn}I_{ip,nj}$ and the diagram E_{pn}. In that case we impose the condition $[npij]$ of the introduction. $B_{jipn}I_{ip,nj}$ reverses all the inequalities of the condition and is an automorph if all the equalities hold. We next show that if one of the conditions 1) to 5) holds the application of the transformation to any form whose every $[\cdots]$ and $\{\cdots\}$ is satisfied yields a form in which one of these is not satisfied unless the transformation has the same effect as some permissible transformation previously considered.

First we show that if the transformation is applied to a diagram E_{pn} in which $[jip]$ holds and $jn \neq ip$, $[jip]'$ holds if and only if $mn + jn < in$ unless the application of the transformation has the same effect as some permissible transformation previously considered. $B_{jipn}I_{ip,nj}$ takes $[jip]_1$ into $mn + jn\ \alpha'\ in$ and hence if $[jip]$ is to hold α' is $\leqq$. If α' is $=$, $[jip]'_2$ is $mn\ \beta'\ pn$, $pn\ \beta''\ jn + ij$ and, if β' is not $=$ but β'' is, $jp\ \beta'''\ pm + ip$. Hence β' or β'' is $=$. In the former case, $A_{jip}I_{jp,mn}$ is an automorph of the resulting form and $B_{jipn}I_{ip,jn}A_{jip}I_{jp,mn} = B_{jipn}A_{npi}I_{ip,jn}I_{jp,mn} = A_{jip}I_x$ concludes the proof. The latter case we next consider.

Second consider 2) with all the $[\cdots]$ and $\{\cdots\}$ holding. We prove $[ijn]'$ holds if and only if $jp \leqq jm$. After the application of $B_{jipn}I_{ip,jn}$ $[ijn]'$ is $jp\ \gamma'\ jm$ and hence if $[ijn]'$ is to hold, γ' is $\leqq$. Hence, in the case postponed above, if $mn + jn = in$ and, from $[jip]$: $jm \leqq pm + ip$ we have $jp \leqq pm + ip$ and β''' is $\leqq$ while β' is $\geqq$. The case $\beta' =$ has already been considered. If β''' and β'' are $=$, $A_{jip}I_{jp}$ is an automorph for the transformed form and we proceed as above. It remains to prove $[ijn]'$ holds if $jp = jm$. This follows since $[ijn]$ implies $im \leqq nm + ip$ and $[ijn]'_2$ holds.

Third consider 3) with all $[\cdots]$ and $\{\cdots\}$ holding. Note that the transformation takes the diagram F_{pn} into H_{pn} and if $mn + jn < in$ and 2) does not hold we prove $[ijp]'$ holds if and only if $jp + pm < in$ unless $B_{jipn}I_{ip,jn}$ has the same effect as some transformation previously provided for. $[ijp]'$ is $jp + pm\ \delta'\ in$ and we need inquire further only if δ' is $=$. If $[ijp]'_2$ holds then $jp + ij\ \epsilon'\ pn\ \epsilon''\ pm$. If ϵ'' is $=$ then $[jip]$ implies $jm = jn = ip$ and 2) implies $jp = jm$. $\delta' =$ implies $jp = in$ and $mn + jn \geqq in$ which cannot be. If ϵ' is $=$, $A_{ijp}I_{ip,mn}$

is an automorph for the transformed form and $B_{jipn}I_{ip,jn}A_{ijp}I_{ip,mn} = A_{jip}A_{ipn}A_{pni}I_x = A_{jip}A_{nip}I_y = I_{-i,n}A_{mip}I_y$ concludes the proof.

Fourth, consider 4) with all $[\cdots]$ and $\{\cdots\}$ holding. The transformation takes H_{pn} into F_{pn} and we prove if $mn + jn < in$, then $[jim]'$ and $\{jin\}'$ hold if and only if 4) is denied unless it has the same effect as some previous transformation. $\{jin\}'$ is $im + mn\ \lambda'\ jm$ and hence is denied if λ' is $>$. If λ' is $=$ and $\{jin\}'_2$ is to hold, $mn\ \mu'\ pn$, $pn\ \mu''\ im + ij$ and hence one of the equalities holds. Then $I_{-i,n}A_{mip}I_x$ is an automorph and $B_{jipn}I_{ip,jn}I_{-i,n}A_{mip}I_x = B_{jipn}I_{-p,j}A_{mpi}I_y = A_{jip}A_{ipn}A_{ipn}A_{ipj}A_{ipm}A_{mpi}I_y = A_{jip}A_{ipj}I_y = A_{ijp}I_x$ complete the argument for $\{jin\}'$. If then $im + mn < jm$, $[jim]'$ is $im + jn < in + jm$ for $mn + jn < in$ implies $jn < in$.

Finally we prove that if $jm = ip$, $mn = 0$ and $[ijm]$ holds, the transformation takes the form into one in which $[ijm]'$ does not hold unless it has the same effect as a transformation previously provided for. $[ijm]$ implies $in \leqq pm + ip$ while $[ijm]'$ implies $pm + ip \leqq in$ and hence $pm + ip = in$. In this case $A_{ijm}I_{im,pn}$ is an automorph and to complete the proof we use $A_{ijm}I_{im,pn}B_{jipn}I_{ip,jn} = A_{ijm}B_{jmnp}I_x = A_{mnp}I_{ij,pm}$.

From the above discussion, we have the following cases in which $[npij]$ does not apply:

C_1: $in = ip$, for then $mn + jn \geqq in$.

C_2: $im = jp = ip$, for $[ijp]$ implies $in \leqq pm + jp$ and 3) holds.

C_3: If $jp = ip$ and $[npij]$ applies, then $[npji]$ does not for $im + mn < jm$ implies $jm + mn > im$.

C_4: If $jn = ip$ and $[npij]$ applies, then $[pnji]$ does not for $[jip]$ implies $jm \leqq pm + ip$ while $[pnji]$ and $[npij]$ applying imply $mp + ip < jp \leqq jm$.

Next consider the transformation $I_{-jp,m}B_{ipnj}$ which is permissible for the diagram G_{pn} and makes the following changes:

interchanges pj and $pm + pi$, that is, $pj + ij$ and pm,

and interchanges in and $jn + nm$, that is, $in + ij$ and nm.

We impose the condition $\{npij\}$ of the introduction. To see that this may be imposed, note that the transformation reverses the inequality of $\{npij\}_1$ and in the first part of $\{npij\}_2$ but leaves all others, whereas $jp + ij = pm$ implies that the transformation is an automorph. If α' and α'' are $=$, $I_{-jp,m}B_{ipnj}I_{pn,ij}$ is an automorph.

We next prove that if 1) and all conditions $[\cdots]$ and $\{\cdots\}$ hold, $[jip]'$ and $[ijn]'$ hold and conversely if all $[\cdots]$ and $\{\cdots\}$ hold and if after the transformation $[jip]'$ and $[ijn]'$ hold, then 1) must hold unless the transformation has the same effect on the form as a transformation previously considered. $[jip]'$ and $[ijn]'$ are $jm \leqq jp$ and $im \leqq in$ respectively which proves the first part of the above statement. We will prove the second by considering the case when $jm = jp$. Then $[jip]'_2$ implies $jm + ij\ \alpha'\ pn$, $pn = jn + ij$, $im\ \alpha''\ jn + nm$. α' is $\geqq$

while $[ijn]$ implies α'' is $\leqq$. Hence α' or α'' is $=$ and $A_{jip}I_x$ is an automorph for some x. $I_{-jp,m}B_{ipnj}A_{jip}I_x = B_{npij}I_y$ and hence the application of the transformation has the same effect as a previous transformation. We proceed similarly if $im = in$.

Finally, if $im = ip$ and all $[\cdots]$ and $\{\cdots\}$ hold we prove that if 1) and 2) hold, then after the transformation $\{jin\}'$ and $[jim]'$ hold and conversely that if 1) holds and if $\{jin\}$ and $\{jin\}'$ hold, then 2) holds unless the application of the transformation has the same effect on the form as a transformation previously considered. $\{jin\}'$ is $pm + jm \leqq in$ and hence is satisfied if 2) holds and in that case, by A2 below, $[jim]'$ also holds. Conversely if $\{jin\}'$ holds 2) must hold unless $pm + jm = in$. In that case $\{jin\}'_2$ implies $jm + ij\ \beta'\ pn\ \beta''\ pm$ and hence β' or β'' is $=$. If β'' is $=$, $[jip]$ says $jm \leqq ip$. Hence in either case β' is $=$ and $I_{-i,n}A_{pim}I_{jn}$ is an automorph for the resulting form. $I_{-jp,m}B_{ipnj}$ $I_{-i,n}\ A_{pim}I_{jn} = I_{-i,n}A_{pim}I_x$ completes the proof.

Note that 5.2) and 6.1) show that $[npij]$ and $\{npij\}$ are not simultaneously applicable.

We shall make use of the following cases in which various conditions automatically hold.

A1. $[npij]$ automatically holds if it applies and $jm + ij = 0$, for $[ijm]$ implies $in \leqq pm + mn + jm$ and if the $=$ holds, $pm \leqq in + ij$. If $pm = in + ij$, then $mn = 0$ and the condition does not apply.

A2. If in the diagram J_{pn} $\{jin\}$ holds, then $[jim]$ must hold for $\{jin\}$ is $jp + jm + ij \leqq mn + ip$ and $[jim]$ is $jp + jn + ij\ \alpha'\ pm + mn + im$. And if α' is $\geqq$ we would have $pm + mn + im \leqq jp + jn + ij \leqq jp + jm + ij \leqq mn + ip$ and hence $pm = 0$, $jm = jn$, $jp + ij = mn$ which implies that α' is $=$ and $[jim]_2$ holds.

A3. In diagram J, $[ijn]$ automatically holds for $im + ij \leqq mn + pn$ and if the equality holds, $im + ij = pn$.

A4. If in L_{pm}, $[jim]$ holds, $\{jin\}$ must hold for $[jim]$ is $jp + jn + ij \leqq mn + im$ and $\{jin\}$ is $jp + jm + ij\ \alpha'\ pn + mn + ip$. If α' is $\geqq$ we have $jp + jn + ij \leqq mn + im \leqq pn + mn + ip \leqq jp + jm + ij$ and therefore α' is $=$ and $\{jin\}_2$ holds.

A5. If in L_{mn}, $[jim]$ holds then $[jip]$ must hold.

A6. In N $\{jin\}$ and $\{ijn\}$ automatically hold.

A7. In L_{pn} $[jip]$ and $\{jin\}$ imply $[jim]$.

A8. In M'_{mn} obtained from M_{mn} by interchanging p and m, the satisfaction of $\{ijp\}$ implies that $\{npij\}$ is not applicable for the former is $in \leqq mp + jm$.

A9. $[npij]$ is not applicable, from C_1, if ni occurs in the diagram. If line jm occurs, none of jn, jp, in, mn occur. Furthermore, if lines jp and im occur and $[ijp]$ holds, the condition is not applicable for $[ijp]$ implies $in \leqq pm + jp$ which is condition 3). Hence, using lemma 2 we see that, of the diagrams we need consider, in only the following can $[npij]$ apply: E_{pn}, F_{pn}, G_{pn}, G'_{pm}, H_{pn}, J_{pn}, L_{pn}, L''_{pm}, E_{pnm}, H_{pnm}, G'_{pmn} where the primed diagrams are defined below.

It is all the more true that no condition $[\cdots]$ is applicable to a diagram containing beside ij more than three non-zero lines and in which the $[\cdots]$ and $\{\cdots\}$ conditions hold.

A10. $\{npij\}$ is clearly not applicableif either of the lines jp or ni occur. If im and jm occur the condition is not applicable if $\{ijp\}$ holds from A8. Hence no condition $\{\cdots\}$ is applicable to a diagram containing beside ij more than three non-zero lines.

A11. The condition (pnm) applies only to E_{pnm}, H_{pnm}, G_{pmn} as may be seen from the discussion above for this condition.

Consider the diagrams E_{pn} and G. Since $A_{jip}I_{jp}$ takes one into the other, we may assume in the former that $[jip]$ holds and in the latter $[ijn]$ holds. If in the latter $[jip]$ does not hold, the transformation will take G into E_{pn} with $[jip]'$ holding. Hence we have either G in which $[jip]$ and $[ijn]$ hold or E_{pn} in which $[jip]$ holds. If $[npij]$ applies but does not hold, use $B_{jipn}I_{ip,jn}$ as above which does not alter the diagram.

Consider the three diagrams F_{pn}, L and H'_{pn} where the last is H_{pn} with p and m interchanged. $A_{jip}I_{jp,mn}$ interchanges the first two diagrams while $A_{jim}I_{jm,pn}$ leaves them unaltered. $A_{ijm}I_{im,pn}$ interchanges the second and the third diagrams. $I_{-i,n}A_{mip}I_{jn}$ interchanges the first and L' where L' is obtained from L by interchanging p and m. Thus in these diagrams we may assume that $[jip]$, $[jim]$ and $\{jin\}$ hold when they apply.

First, the only other condition which can apply to F_{pn} is $[npij]$ and by the discussion of that condition, if it applies but is not satisfied, the transformation $B_{jipn}I_{ip,nj}$ takes the form into one in which all $[\cdots]$ and $\{\cdots\}$ hold as well as $[npij]'$.

Second, if in L, $[jim]$ but not $[ijm]$ holds, use $A_{ijm}I_{im,pn}$ which converts the diagram into H'_{pn} in which $[ijm]'$ holds. It can be shown that $[jim]'$ holds also. In H'_{pn} $[npij]$ is taken care of by previous theory as mentioned in the first case.

Third, if in H_{pn}, $[ijp]$ is satisfied but $[jip]$ is not, apply $B_{jipn}I_{ip,jn}$ which takes the diagram into F_{pn}. This, in turn may be taken into F_{pn} or L in which $[jip]$, $[jim]$ and $\{jin\}$ are satisfied and then proceed as above. $[nipj]$ is taken care of by previous theory.

Consider the diagrams containing F_{pm} and J. $A_{jip}I_{jp,mn}$ interchanges F_{pm} and J and $I_{-i,n}A_{mip}I_{jn}$ leaves them unaltered. Hence we assume $[jip]$, $[jim]$ and $\{jin\}$ hold in both. In the first, $[mpij]$ and $[pmij]$ do not apply by C_1 and in the second $[ijn]$ automatically holds by A3.

Consider the diagrams G'_{pm} and E'_{pnm} the former being obtained from G_{pm} by interchanging m and n and the latter from E_{pnm} by interchanging i and j as well as p and m. $A_{jip}I_{jp,mn}$ leaves G'_{pm} while $A_{ijm}I_{im,pn}$ interchanges the diagrams. If $[jip]$ holds in G'_{pm} but $[ijm]$ does not use the second transformation to yield E'_{pnm} in which $[ijm]$ holds. Hence in both diagrams $[ijm]$ may be assumed to hold as well as $[jip]$ in the first. In $G'_{pm}[npij]$ automatically holds by A1 if it applies. If $[nipj]$ is denied but applicable in E_{pnm} and $[jip]$ holds, the former implies $jm \neq ip$ and $pm + mn + ip \leqq in + jm + ij$ which is, after the application of $B_{jipn}I_{ip,nj}$, the condition $[ijm]'_1$. If the equality holds,

$pm \geqq in + ij$ and $[ijm]_2'$ is $ip' + ij' < mn' \leqq in' + ij'$, that is, $jm + ij \leqq pm$. If $jm + ij = pm$, then $in = ip$ which contradicts the applicability of $[npij]$. $[jip]'$ holds if $[npij]$ is applicable. If now in E_{pnm}, (pnm) is applicable and denied, the transformation A_{pnm} may be used with $[jip]$ still holding. If then $[npij]$ is denied proceed as above.

Consider any diagram containing G_{pn}; we prove that the conditions associated with that part of the diagram which is G_{pn} can be satisfied simultaneously. Suppose $[ijn]$ holds but $[jip]$ does not. Then $im \leqq mn + jn$ and $jm \geqq pm + ip$. First if $in > mn + ip$ and $jp \leqq jm$, the theory connected with $[npij]$ shows that $B_{jipn}I_{ip,jn}$ yields a diagram in which $[jip]'$ and $[ijn]'$ hold. Second, if $in > im$ (this is implied by $in > mn + ip$) and $jp > jm$, use $I_{-jp,m}B_{ipnj}$. Third, if $in \leqq mn + ip$, $A_{jip}I_{jp}$ leaves the diagram and takes $[ijn]$ into $in \leqq mn + ip$. If the equality holds $[ijn]_2'$ is $im' + ij'\ \alpha'\ pn' = ip' + ij'$ and $jm'\ \beta'\ jp'$, that is, $in + ij \geqq pn$ and $pm + ip\ \beta'\ jp$ which holds unless β' is $<$ and $in \neq ip$. In this case application of $B_{jipn}I_{ip,jn}$ or $I_{-jp,m}B_{ipnj}$ instead of $A_{jip}I_{jp}$ yields a diagram in which $[ijn]'$ holds and $[jip]'$ holds since $nm \geqq pn = jn + ij$, $jp > pm + ip$. Hence we have a diagram G_{pn} in which $[jip]$ and $[ijn]$ hold. The previous theory takes care of the other conditions.

Consider any diagram containing G'_{pmn} obtained from G_{pmn} by interchanging m and n. $A_{jip}I_{jp,mn}$ and $A_{ijm}I_{im,pn}$ leave the portion of the diagram containing G'_{pmn}. Assume $[ijm]$ holds but $[jip]$ does not, that is, $in \leqq pm + jm < jn$. $A_{jip}I_{jp,mn}$ gives as $[ijm]'$, $in < jn$. A_{pnm}, if necessary, may be applied to make (pnm) hold without disturbing the other conditions.

Consider any diagram containing J_{pm}. $[ijn]$ is automatically satisfied. The transformations for $[jip]$, $[jim]$ and $\{jin\}$ leave that portion of the diagram and no other conditions apply.

Consider diagrams containing J_{pn}, L_{pm} and M. Write L'_{pm} and M' for the diagrams obtained from L_{pm} and M respectively by interchanging p and m. $A_{jip}I_{jp,mn}$ interchanges J_{pn} and L_{pm} as well as L'_{pm} and M. $A_{jim}I_{jm,pn}$ interchanges L_{pm} and M' while $I_{-i,n}A_{mip}I_{jn}$ interchanges J_{pn} and M. Note that, if in J_{pn}, $[jim]$ does not hold $\{jin\}$ does not from A2 and we use the third transformation above. Also, if in L_{pm}, $[jim]$ holds, $\{jin\}$ holds by A4. Hence in the above diagrams we may assume $[jip]$, $[jim]$ and $\{jin\}$ hold. Note that though $I_{-i,n}A_{mip}I_{jn}$ was used in fixing $[jip]$ and $[jim]$ in diagram F it is equal to $A_{jip}I_{jp,mn}A_{jim}I_x$ and hence will yield no new diagrams.

First, in J_{pn} the previous theory takes care of $[npij]$, $[pnji]$ and $\{npij\}$ when they apply.

Second, call L''_{pm} the diagram obtained from L_{pm} by the permutation (ij, mpn). Suppose $[ijp]$ and $[ijn]$ hold but $[jip]$ does not. If $mn + jn < in$ conditions 2) and 4) in $[npij]$ do not hold and hence $B_{jipn}I_{ip,jn}$ yields a diagram in which all the $[\cdots]$ and $\{\cdots\}$ hold. On the other hand, if $mn + jn \geqq in$, use $A_{jip}I_{jp}$ which leaves the diagram and gives a form for which the conditions $[\cdots]$ hold and for which A4 implies that $\{ijm\}$ holds. The conditions $[\cdots]$ are covered by previous theory and the conditions $\{\cdots\}$ do not apply.

Third, if in M, $[ijp]$, $[ijn]$ and $\{ijm\}$ hold, note that $[jim]$ holds from A3.

Suppose that $[jip]$ does not hold. If $pm \leqq nm$, use $A_{jip}I_{jp,mn}$ to fix all conditions. If $pm > mn$, $\{jin\}$ is denied and use $I_{-i,n}A_{mip}I_{jn}$ which changes the diagram to J_{pn} with $\{jin\}'$ and $[jip]'$ holding. The conditions $[jim]'$ and $[ijn]'$ hold automatically and other conditions are covered by previous theory. Note that if $pm > mn$ and $\{jin\}$ holds, then $[jip]$ does.

Consider L'_{mn} and N where L'_{mn} is obtained from L_{mn} by interchanging p and m. $A_{jip}I_{jp,mn}$ interchanges the diagrams while $I_{-i,n}A_{mip}I_{jn}$ and $A_{ijp}I_{ip,mn}$ leave the former. Hence we may suppose $[ijp]$ and $[jip]$ hold in the former and $[ijm]$ and $[jim]$ in the latter. First consider L'_{mn} with $[ijp]$ and $[jip]$ holding. $[jim]$ holds by A5. If $\{jin\}$ does not hold, use $I_{-i,n}A_{mip}I_{jn}$ which fixes all the conditions $[\cdots]$ and $\{\cdots\}$. Second, if we have N with $[ijm]$ and $[jim]$ holding, from symmetry we take $in \leqq jn$. Then if $[jip]$ holds, $[ijp]$ will also. If $[jip]$ does not hold, use $A_{jip}I_{jp,mn}$ to fix all $[\cdots]$. The $\{\cdots\}$ in N conditions hold by A6.

The transformations which fix the diagram K in K_{mn} and P interchange or leave the diagrams and hence in each the conditions associated with K may be assumed to hold. These diagrams are then easily dealt with.

Consider L_{pn} and H'_{pnm} where the latter is obtained from H_{pnm} by interchanging p and m. $A_{jim}I_{jm,pn}$, $A_{jip}I_{jp,mn}$ and $I_{-i,n}A_{mip}I_{jn,pm}$ leave L_{pn} and $A_{ijm}I_{im,pn}$ interchanges L_{pn} and H'_{pnm}. Hence we may assume that in L_{pn}, $[jip]$, $\{jin\}$ and hence, by A7 $[jim]$ also holds. If $[ijm]$ does not use $A_{ijm}I_{im,pn}$ which takes L_{pn} into H'_{pnm} with $[ijm]'$ and $[jim]'$ holding. If one of $[ijm]$, $[jim]$ in H'_{pnm} fails to hold the diagram may be taken into L_{pn} with $[ijm]$ holding. Hence, in both these diagrams we can make the conditions $[\cdots]$ and $\{\cdots\}$ to hold. $[npij]$ automatically holds if it applies in L_{pn}. If in H_{pnm}, $[npij]$ applies but does not hold, $B_{jipn}I_{ip,nj}$ will take the diagram into L_{pn} with all conditions holding. If in $H_{pnm}(pnm)$ is applicable, use A_{pnm} to fix all $[\cdots]$ conditions. If, after this, $[npij]$ is denied, proceed as above.

Consider diagrams containing N_{pm} and M_{pn}. $A_{jip}I_{jp,mn}$ interchanges the diagrams. First, by previous theory, we may assume that $[ijm]$ and $[jim]$ in N_{pm} hold. If $[jip]$ does not hold, we apply $A_{jip}I_{jp,mn}$ to take N_{pm} into M_{pn} in which by the theory for N (assuming from symmetry that $in \leqq jn$) $[jim]'$, $[jip]'$, $\{jin\}'$ and $[ijp]'$ hold. $[ijn]'$ and $\{ijm\}'$ hold automatically. Second, since $A_{ijp}I_{ip,mn}$ leaves the diagram M_{pn} we may assume $[ijp]$ holds. From A3 and A4 $[jim]$, $[ijn]$, $\{ijm\}$ automatically hold. Suppose $[jip]$ does not hold. If $pm \leqq nm$, the theory for M may be used to show that $A_{jip}I_{jp,mn}$ takes M_{pn} into N_{pm} with all applicable conditions holding. On the other hand, if $pm > mn$, $I_{-i,n}A_{mip}I_{jn}$ takes M_{pn} into itself with all applicable conditions holding.

Consider the diagrams containing N_{pn}. Since $A_{jip}I_{jp,mn}$ and $A_{ijp}I_{ip,mn}$ leave the diagram N_{pn} we may assume that $[ijp]$ and $[jip]$ hold. Then $\{ijn\}$ and $\{jin\}$ automatically hold from A6 and hence $[ijm]$ holds from A7 and $[jim]$ from A5.

Consider the diagram N_{pnm}. The theory for G'_{pmn} together with A5 dispose of this case.

Consider the diagram M'_{mn} obtained from M_{mn} by interchanging p and m.

The transformation $I_{-jp,m}B_{ipnj}$ leaves the diagram and may be used if necessary to make $jp + in + ij \leqq pm + nm + ip$. $[jip]$ and $[ijn]$ automatically hold. If $jp > mn + ip$, then $in < pm + ip$ and application of $I_{-i,n}A_{mip}I_{jn,pm}$ fixes all conditions. The case when $in > pm + ip$ is dealt with symmetrically. And finally, if $jp \leqq mn + ip$, $in \leqq pm + ip$, $\{jin\}$ and $\{ijp\}$ hold and hence also the other conditions.

The diagrams P_{mn}, P_{pm}, Q_{pn} and Q_{mn} are easily dealt with if we first fix the conditions of diagram K which they contain. $A_{ijp}I_{ip,mn}$ takes Q into a diagram like P_{mn}. In R all conditions $[\cdots]$ and $\{\cdots\}$ automatically hold and no others apply. It remains to consider all diagrams containing N_{pnm}. These are easily disposed of.

We now show that if the transformation $C_{ijm,pn}I_{ijm}$ is permissible, its application to a diagram in which all applicable conditions above hold, yields a form in which at least one applicable condition does not hold unless the transformation is an automorph. The transformation makes the following changes in coefficients $jn' = ip$, $ip' = jm$, $jp' = in$, $nm' = ij + nj$, $in' = jm$, $mp' = ij + jp$. If the transformation is permissible $ij + jm = mp = nm = 0$. $[ijm]$ implies $ip = in = mj$ and $\{jim\}$ implies $jp = jn = ip$. Then, the transformation is an automorph.

Similarly $I_{-i,m}C_{mji,pn}I_{imj}$ makes the following changes in coefficients: $jn' = mp + ip$, $jm' = in$, $jp' = nm + ni$, $ni' = nj$, $mn' = 0$, $ip' = jp$, $mp' = 0$ and is permissible if $in = ip = jm = -ij$. $[ijm]'$ implies $ip' = in' = jm'$ and $\{jim\}'$ implies $jp' = jn' = ip'$. Therefore $jp = jn = in = nm + ni = mp + ip = ip = jm$ and hence $mp = mn = 0$ and the transformation is an automorph.

4. Symmetric conditions for $a_{ij} > 0$, $a_{nm} > 0$ and all other coefficients < 0

Throughout this section ij in some order and mn in some order are the subscripts of the two positive coefficients. The first two conditions given in the introduction arise from the earlier paper.

The effect of $A_{jip}I_{jp,mn}$ is given in the previous section and hence we use the former condition $[jip]$ when $ji + ip = 0$ with the exceptions noted. These exceptions are dealt with later. Recall that this transformation interchanges the lines jn and mp, jm and np in any diagram. Note that if $jn + ij + mn = mn + jm + ij = 0$ the condition $[jip]$ automatically holds.

If $ij + jn + mn = 0$, B_{ijnm} is permissible and we accordingly list the changes in coefficients resulting from this transformation and one other:

B_{ijnm} :	jn'	ij'	jm'	jp'	in'	nm'	np'	im'
$B_{ijnm}I_{jn,im}$:	jn'	nm'	in'	np'	jm'	ij'	jp'	im'
Original form:	jn	nm	$jp - nm$	$ij + jm$	$np - ij$	ij	$in + nm$	im

B_{ijnm} :	ip'	mp'
$B_{ijnm}I_{jn,im}$:	mp'	ip'
Original form:	$ij - nm + ip$	$mp + nm - ij$

We accordingly set up the condition c)3. of the introduction. To establish its validity note that $B_{ijnm}I_{jn,im}$ reverses the inequality of the first line and α'

but not α''. It reverses α''' while if α' and α''' are $=$ it is an automorph. It leaves α^{IV} while if α'' and α^{IV} are $=$, B_{ijnm} is an automorph. In the third line of $[ijnm]_3$ if α''' or α^{IV} is $=$, one or more of the transformations just mentioned is an automorph. Note that if m and i, j and n are interchanged the condition remains unaltered.

In this section our diagrams contain the lines ij and mn, the line jn, for instance, if $ij + jn + nm = 0$ and the line jp, for instance, if $ij + jp = 0$.

By means of the following considerations we eliminate from consideration certain diagrams. We call ip, jp, np, mp *loose* lines and jm, jn, im, *in joined* lines. First if just two loose lines ip and pn appear, use $A_{jip}I_{jp,mn}$ unless one of jn, jm appears or $A_{mnp}I_{ij,mp}$ unless one of the lines im, jm appears. Hence if the number of joined lines is less than two, the number of loose lines in the diagram can be decreased unless jm appears. Second if just two loose lines ip and jp appear and jn appears use B_{ijnm} unless one of in, jm occurs. Third, if ip, np, mp occur, $A_{jip}I_{jp,mn}$ decreases the number of loose lines unless both jn and jm occur. Finally, if all four loose lines occur in the diagram it may be taken into one containing fewer loose lines unless all four joined lines occur. Hence *every diagram may be taken by a permissible transformation into one of the following types*:

1. Having one or less loose lines,
2. Having only the lines ip, pn, jm (diagram H_8),
3. Having only the lines ip and jp (diagram G_6),
4. Having only the lines ip, pn, mp, jn, jm (diagram J_5'),
5. Having at least as many joined lines as loose lines.

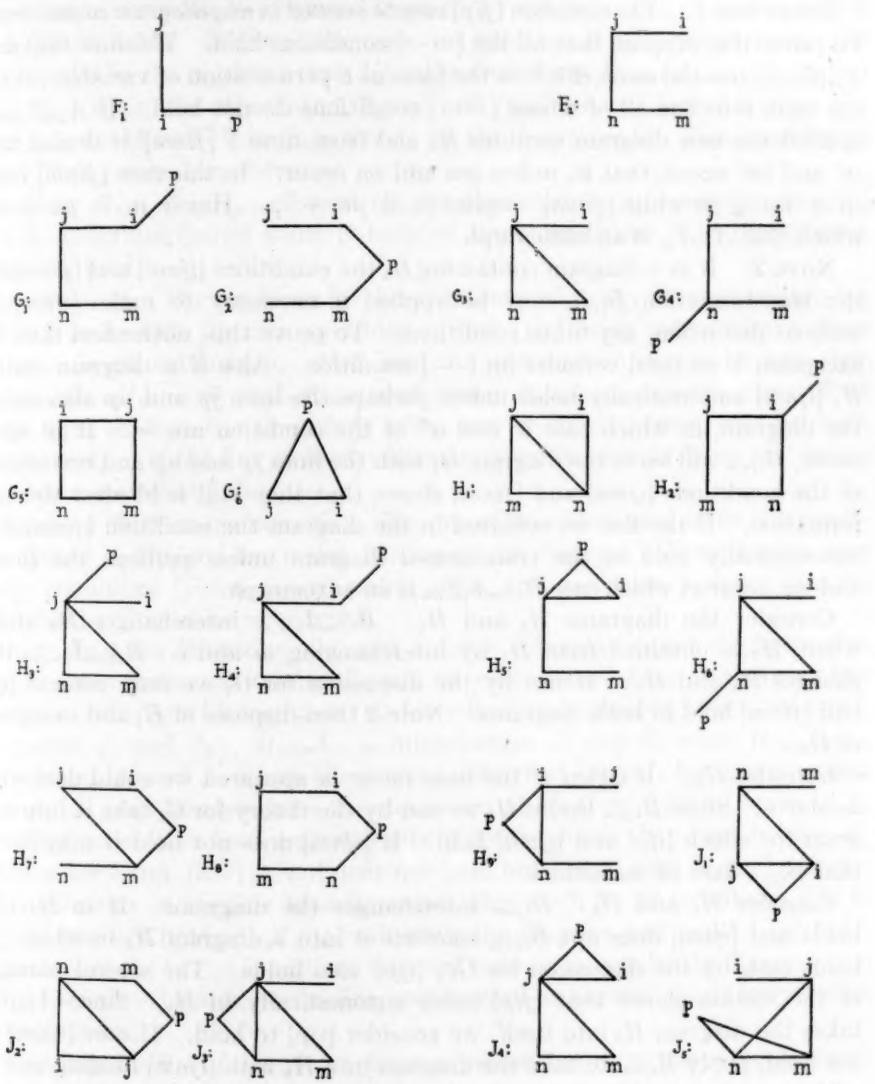
We accordingly display the diagrams of set III to be considered and proceed to consider all diagrams of these types. Note that F_1 and F_2 have already been considered while G_6 has been considered in the previous section.

We first find a preliminary result. Suppose we have a diagram containing G_1 and that in $[ijnm]_1$ and $[jip]_1$ the equalities hold while $[jip]$ is satisfied but $[ijnm]$ is not. We show that the application of B_{ijnm} results in the satisfaction of $[jip]'$ as well as $[ijnm]'$. By hypothesis $jm + ij\ \alpha'\ np$, $pn\ \alpha''\ jn + ij$ and $jm + ij\ \alpha'\ np$, $in + ij\ \alpha'''\ np$. $[jip]_2'$ is $jp\ \beta'\ in + nm$ and $in \geqq jn$. Hence $\alpha' + \alpha''$ is $\geqq$ and $jp + np + jn = jm + in$ implies $\beta' = \alpha'$. If β' is $=$ $[jip]_2'$ is automatically satisfied. If it is not there is doubt only when $in = jn$. Then $[jip]_3'$ is $im\ \delta'\ np - ij$ and $im \geqq in \geqq np - ij$ completes the proof.

Consider the diagrams G_1 and G_2. $A_{jip}I_{jp,mn}$ interchanges the diagrams. If in G_2 $[jip]$ or $[nmp]$ is denied, $A_{jip}I_{jp,mn}$ or the corresponding transformation will take it into G_1 with $[jip]$ satisfied. If in G_1 $[jip]$ is denied we may send it into G_2 and if necessary back into G_1. Hence we need only consider G_1 with $[jip]$ satisfied and $[ijnm]$ not. B_{ijnm} leaves the diagram and $[jip]'$ is $jp + jn\ \alpha'\ mp + nm + in + ip$ and α' is $\leqq$ from $[jip]$ and $[ijnm]$, the equality holding only if the equality $[jip]_1$ and $[ijnm]_1$ holds in which case the result above completes the proof.

Consider the diagrams G_3 and G_4. $B_{ijnm}I_{jn,im}$ interchanges the two dia-

grams. If in G_3 $[ijnm]$ does not hold, $B_{ijnm}I_{jn,im}$ takes the diagram into G_4 whose condition $[ijnm]$ holds. If in G_4 $[ijnm]$ does not hold, the diagram may be taken into G_3 and dealt with as above. If in G_4, $[ijnm]$ holds and $[mnp]$ does not it may be shown that application of $A_{mnp}I_{mp,ij}$ leaves the diagram and yields a form for which $[ijnm]$ and $[mnp]$ both hold.



Consider the diagram G_5. $B_{jinm}I_{in,jm}$ does not alter it. If then we assume that $[ijmn]$ holds but $[jinm]$ does not it may be shown that $B_{jinm}I_{in,jm}$ yields a form for which both conditions hold.

Before proceeding with the diagrams H we prove some more general results.

NOTE 1. If a diagram contains H_9, $[jinm]$ is denied unless the lines im and jn appear for $ip + np + in\ \alpha'\ im + jn$ implies and is implied by $0\ \alpha'\ im + jn + 2mn + 2ij$. If these two lines do appear the equalities in $[jinm]_2$ hold. If, on the other hand, they do not both appear, B_{jinm} will decrease the number of loose lines.

EXCEPTION 1. *The condition* $[jip]$ *may be omitted in any diagram containing* H_2. To prove this suppose that all the $[\cdots\cdot]$ conditions hold. We show that unless $A_{jip}I_{jp,mn}$ has the same effect on the form as a permutation of variables, it takes the form into one all of whose $[\cdots\cdot]$ conditions do not hold. If $A_{jip}I_{jp,mn}$ is applied the new diagram contains H_9 and from note 1 $[jinm]'$ is denied unless jn' and im' occur, that is, unless pm and im occur. In this case $[jinm]$ implies $ip + np \geqq jn$ while $[jimn]'$ implies $jn \geqq pn + ip$. Hence $jn = pn + ip$ in which case $A_{jip}I_{jp}$ is an automorph.

NOTE 2. If in a diagram containing H_1 the conditions $[ijmn]$ and $[jinm]$ hold, the transformation B_{ijnm} may be applied if necessary to make $[ijnm]$ hold without disturbing any other conditions. To prove this, notice first that, from exception 1 we need consider no $[\cdots]$ condition. Also if a diagram contains H_1 $[ijnm]$ automatically holds unless perhaps the lines jp and np also occur in the diagram, in which case α' and α'' of the condition are $=$. If jp and np occur, B_{ijnm} will leave the diagram H_1 with the lines jp and np and investigation of the conditions $[ijmn]$ and $[jinm]$ shows that they still hold after the transformation. If the line im occurred in the diagram the condition $[jimn]$ would automatically hold in the transformed diagram unless perhaps the lines ip and mp occur in which case $B_{ijnm}I_{jn,im}$ is an automorph.

Consider the diagrams H_1 and H_2. $B_{ijmn}I_{jm,in}$ interchanges H_1 and H_2' where H_2' is obtained from H_2 by interchanging m and i. $B_{jinm}I_{jm,in}$ interchanges H_1 and H_2. Hence by the discussion for G_5 we may assume $[ijmn]$ and $[jinm]$ hold in both diagrams. Note 2 then disposes of H_1 and exception 1 of H_2.

Consider H_3. If either of the lines im or in appeared we could deal with it as above. Since B_{ijnm} leaves H_3 we can by the theory for G_4 take it into a diagram for which $[ijp]$ and $[ijmn]$ hold. If $[ijnm]$ does not hold it may be seen that B_{ijnm} fixes all conditions.

Consider H_4 and H_5. B_{ijnm} interchanges the diagrams. If in H_4 $[ijmn]$ holds and $[ijnm]$ does not B_{ijnm} converts it into a diagram H_5 in which $[jip]'$ holds and, by the discussion for G_3, $[ijp]'$ also holds. The second paragraph of the section shows that $[jip]$ holds automatically in H_4. Since $A_{ijp}I_{ip,mn}$ takes the diagram H_5 into itself, we consider $[ijp]$ to hold. If now $[ijnm]$ does not hold, apply B_{ijnm} to take the diagram into H_4 with $[ijnm]$ holding and deal with it as above. If, on the other hand, $[ijnm]$ does hold, it may be seen that $[jip]$ also holds.

Consider H_6 and H_7. $B_{ijmn}I_{jm,in}$ interchanges the diagrams while $B_{ijnm}I_{jn,im}$ leaves H_6. We may then consider $[ijmn]$ to hold in both diagrams. If then in H_6, $[ijnm]$ does not hold, it may be seen that application of $B_{ijnm}I_{jn,im}$

will leave the diagram and fix conditions $[ijnm]$ and $[mnp]$. If in H_6, $[ijnm]$ and $[mnp]$ but not $[ijmn]$ hold, it may be seen that application of $B_{ijmn}I_{jm,in}$ changes the diagram into H_7 where all conditions hold. If in H_6 $[ijnm]$ and $[ijmn]$ but not $[mnp]$ hold, it may be seen with the help of the theory for G_4 that $A_{mnp}I_{mp,ij}$ will take the diagram into a different one with all conditions holding. It remains to consider H_7' obtained from H_7 by interchanging m and n and in which $[ijnm]$ holds. If then $[jip]$ does not hold use $A_{jip}I_{jp}$ which leaves the diagram. If in the resulting form $[ijnm]$ is denied use $B_{ijnm}I_{jn,im}$ and proceed as above. If on the other hand, $[ijnm]$ and $[jip]$ but not $[mnp]$ hold in H_7', $A_{mnp}I_{mp,ij}$ will take H_7' into H_6 in which, by the theory for G_4, $[ijnm]$ and $[mnp]$ hold. Then proceed as above.

Consider H_8. $A_{jip}I_{jp,mn}$ leaves the diagram and may be seen to fix $[jip]$ without disturbing $[mnp]$ while if both of these conditions hold $[ijmn]$ must necessarily hold.

EXCEPTION 2. The condition $[\cdots]$ may be omitted in any diagram containing J_5' which is J_5 with the line ip added. First $[jip]$ holds automatically since $A_{jip}I_{jp,mn}$ is an automorph. Also from the introduction it may be seen that $A_{jip}I_{jp,mn}B_{ijmn}I_{in,jm} = A_{pnm}I_{pm,ij}$ and hence $A_{pnm}I_{pm,ij}$ has the same effect on the form as $B_{ijnm}I_{in,jm}$ which is taken care of above. $[pmn]$ may be dealt with similarly. If the line jp occurs $[ijnm]$ holding implies $jn \geqq in$ and hence exception 1 comes into play.

Suppose a diagram contains G_5. The transformations used to fix the conditions $[jimn]$ and $[ijnm]$ leave this portion of the diagram. Note 2 takes care of any conditions $[jinm]$, $[ijmn]$ which apply and exception 1 shows that no conditions $[\cdots]$ need be considered.

Hence, since the discussion of J_1 and J_2 below also takes care of J_5', it remains to consider diagrams containing G_3 and no other joined lines and two or less loose lines. Those with less than two loose lines have already been considered.

Consider J_1 and J_2. $B_{ijnm}I_{jn,im}$ interchanges J_1 and J_2 while $B_{jinm}I_{jm,in}$ interchanges J_1 and J_2' where the latter is obtained from J_2 by interchanging i and j. It may be seen from the discussion of H_5 that if $[ijnm]$ holds in J_1 then $[jip]$ must also hold and similarly if $[jinm]$ holds then $[ijp]$ holds. If, on the other hand, in J_1 $[ijnm]$ does not hold but $[jinm]$ does it can be shown that application of the first transformation above will fix all conditions. If in J_2 $[ijp]$ does not hold, $A_{ijp}I_{ip,mn}$ will leave the diagram and make $[ijp]'$ hold. If then one of the $[\cdots\cdot]$ conditions does not hold we may take the diagram into J_1 with one $[\cdots\cdot]$ condition holding and proceed as above. In J_2 $[nmp]$ automatically holds.

J_3 is taken by $B_{ijnm}I_{jn,im}$ into a diagram containing three joined lines, which has been dealt with.

Consider J_4. B_{ijnm} and $A_{ijp}I_{ip}$ leave the diagram. If $[ijnm]$ and $[ijmn]$ hold it can be shown that $[ijp]$ holds. By the theory of H_4 and H_5 it remains to consider the case in which $[ijmn]$ and $[ijp]$ but not $[ijnm]$ hold. Application of B_{ijnm} then makes all conditions hold. Note that in this diagram $[jip]$ holds if the other conditions do.

5. Automorphs

The automorphs will be of the two following types: those consisting of permissible transformations A, B, etc., with columns permuted and those which consist merely of permutations of the variables. Any automorph must, of course, leave the diagram of a form unaltered. We consider first those of the first type. Suppose T is a permissible transformation of type A, B, etc., and $TI_x = R$ where R is an automorph. Then $T = RI_x^{-1} = RI_y$. Hence I_y will have the same effect on the form as T. Then, since the application of I_y will deny no symmetric condition, it follows that TI_z must be an automorph for some permutation z which leaves the diagram of T unaltered. Then $(TI_z)(I_z^{-1}I_x) = R$ where $I_z^{-1}I_x$ is an automorph and R is thus a product of an automorph of the first type and an automorph of the second. Hence the only automorphs of the first type which we consider are TI_z where T is a transformation of the type A, B, etc., and z is a permutation of the subscripts which leaves the diagram of T unaltered. These have been found above in the process of setting up the various conditions associated with these transformations. Even in some cases when I_z leaves the diagram, TI_z an automorph implies that T and I_z are. They are the following:

1. If in I a) of the table at the close of this paper, $A_{jip}I_{jp}$ is an automorph, A_{jip} and I_{jp} are; for $A_{jip}I_{jp}$ interchanges im and in and hence is not an automorph unless $im = in$, $jm = pm$, $jn = pn$ in which case A_{jip} and hence I_{jp} are also.

2. If in I a), one of $C_{jip,nm}I_{ij}$ and $C_{jip,nm}I_{ij,nm}$ is an automorph, the other is or, what is an equivalent statement, I_{nm} is. This may be seen from the table.

3. In I b)6. I_{pn} is an automorph.

As noted in the previous paper, the transformation T' on f corresponding to any transformation T on φ dealt with in this paper is obtained by subtracting the bottom element of each column of T from all the elements of that column and then suppressing the last column and last row. For instance, the transformation on f corresponding to the transformation A_{312} in the previous paper is

$$\begin{pmatrix} 1 & 0 & 0 & -1 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix}.$$

We can show as follows that the *determinant of the transformation T' is independent of the subscripts of T*. If T_1 and T_2 are two transformations which differ only in their subscripts, then $T_2 = I_x T_1 I_x^{-1}$ where x is a permutation on the subscripts. Hence by the previous paper $T_2' = I_x' T_1' (I_x')^{-1}$ and hence the determinant of T_2' is equal to the determinant of T_1'. For instance, the transformation T' for $T = A_{312}$ shows that the determinant of each A_{jip} is -1.

Suppose all $a_{ij} \leqq 0$. Then, by the first fundamental theorem of the previous paper all permissible transformations are of the form RT where R is the identity

transformation or a product of permissible transformations Ia 5 or 6 and T is the identity transformation, a matrix formed by permuting the columns of the identity or a permissible transformation A, B, C, D with perhaps its columns permuted. Any permissible transformation $I_{-i,n}$ or $I_{-ij,n}$ is an automorph. *Hence, when $a_{ij} \leqq 0$ for all $i \neq j$, all automorphs are of the form RSU* where R is the identity or a product of automorphs $a5$ or $a6$, S is an automorph of the form $a1$ to $a4$ or the identity, and U is an automorph of type II or the identity.

We proceed to show that any two transformations $a5$ or $a6$ which are simultaneously automorphs are commutative. First suppose $I_{-i,m}$ is an automorph. Making use of corollaries 1 and 2 of the previous paper, we see that the only transformations are the following typical forms: $I_{-j,m}$, $I_{-jn,m}$, $I_{-ij,m}$, $I_{-j,n}$, $I_{-jp,n}$. The first three are obviously commutative with $I_{-i,m}$. Multiplication shows that the fourth and hence the last is. Second, if $I_{-im,n}$ is an automorph, the only other transformations of type $a6$ which can be automorphs are $I_{-ij,n}$ and $I_{-ip,n}$ which are obviously commutative with $I_{-im,n}$.

Hence, all the automorphs of any form may be found as follows from the table: First, if $a_{ij} \leqq 0$ for all ij, put in one set all automorphs of forms 1 to 4 in I a), in the second set all those of forms 5 or 6 and in the third set all those in II. In addition to these automorphs there will be all the automorphs of the first set multiplied on the left by those of the second set in some determined order and these products in turn multiplied on the right by those of the third set. Second, if $a_{ij} > 0$ for some ij the automorphs are those in I, those in II and those of I multiplied on the right by those of II. There will be no duplications except obvious repetitions.

For greater clarification we give two examples.

EXAMPLE 1. If the only zero coefficients are $a_{12} = a_{15} = a_{13} = a_{25} = a_{24} = 0$ and if $a_{23} = a_{14} \neq a_{43} = a_{53} \neq a_{54}$ and all are $\leqq 0$, we draw the diagram of the form, list the permissible transformations and find that the only automorphs in the table are:

$$*I_{-1,4}\ ;\quad *I_{-2,3}\ ;\quad A_{213}I_{45}\ ;\quad *A_{251}I_{21,34}\ ;\quad B_{4215}I_{12,45}\ .$$

In addition we have I, $I_{-1,4}I_{-2,3}$ and the last three above multiplied on the left by $I_{-1,4}$, $I_{-2,3}$ and $I_{-1,4}I_{-2,3}$; then all of these multiplied by $-I$. Hence there are 32 automorphs in all, half of which are of determinant 1. These numbers are halved if we omit $-I$.

EXAMPLE 2. If $a_{ij} > 0$ and all other coefficients are $\leqq 0$; if also the diagram is G_{pn} (see section 3) and $in = im = jm = jp$, $pm = nm$ and $in + ij = pm$. Noting the invariance of the above conditions under an interchange of p and n, i and j, we see that the automorphs in the table are:

$$A_{jip}I_{jp}\ ;\quad A_{ijn}I_{in}\ ;\quad B_{ijnp}I_{ip,jn}\ ;\quad B_{jipn}I_{ip,jn}\ ;\quad I_{-jp,m}B_{ipnj}\ ;\quad I_{ij,pn}\ .$$

We have in addition, the products of the above by $I_{ij,pn}$ on the right—making twelve automorphs not using $-I$. If $-I$ is used there are 24 automorphs. It happens that all are of determinant 1.

A Table of Automorphs[7]

I. First Type

a) $a_{ij} \leqq 0$ for all i, j.

1. If $ji = ip = 0$: $^*A_{jip}$ if $im = in$,
$A_{jip}I_{mn}$ if $jm = jn$ and $pn = pm$,
$^*A_{jip}I_{jp,mn}$ if $jm = pn$ and $jn = pm$.
(If $A_{jip}I_{jp}$ is an automorph I_{jp} and A_{jip} are.)

2. If $ji = ip = pm = 0$: B_{jipm} if $im = in$ and $jp = np$,
$B_{jipm}I_{jm,ip}$ if $im = pn$, $in = jp$, $jn = mn$.

3. If $ji = ip = pn = pm = 0$: $^*C_{jip,nm}I_{ij}$ if $jn = in = jm = im$.

4. If $ij = jn = mn = mp = pi = 0$: $^*D_{ijnmp}$ if[8] $pn = in = jp = jm$,
$^*D_{jnmpi}$ if $im = jm = ni = np$,
$^*D_{nmpij}$ if $jp = pn = jm = im$,
$^*D_{mpijn}$ if $jp = pn = im = in$,
$^*D_{pijnm}$ if $in = jp = jm = im$.

5. If $ij = ip = im = 0$: $^*I_{-i,n}$.

6. If $ip = im = jp = jm = 0$: $I_{-ij,n}$.

b) $a_{ij} > 0$, and other coefficients $\leqq 0$.

1. If $pm = pn = 0$: $^*A_{mpn}$ if $pi = pj$,
$A_{mpn}I_{ij}$ if $mi = mj$ and $ni = nj$,
$^*A_{mpn}I_{ij,mn}$ if $mi = nj$ and $mj = in$.

2. If $ij + ip = 0$: $A_{jip}I_{jp}$ if $im = in$, $jm + ij = pm$ and $jn + ij = pn$,
$^*A_{jip}I_{jp,mn}$ if $jm + ij = pn$ and $jn + ij = pm$.

3. If $-ij = ip = im$: $^*I_{-i,n}A_{mip}I_{jn}$ if $jm = mn + im$ and $jp = pn + ip$,
$I_{-i,n}A_{mip}I_{jn,pm}$ if $jm = pn + ip$ and $jp = mn + im$.

4. If $ij + ip = pn = 0$: $B_{jipn}I_{ip,jn}$ if $im = jp$, $in + ij = pm$ and $mn + ip = jm$.

5. If $ij + ip = ij + jn = pn = 0$: $I_{-pi,m}B_{ipnj}\,I_{ij,pn}$ if $jp + ij = nm$, $pm + ip = in$ and $im = jm$;
$I_{-pj,m}B_{ipnj}$ if $pj = ip + pm$ and $in = nm + jn$.

---

[7] All starred transformations are of determinant -1.

[8] Note that if two of the D's are automorphs, all are. Also since $D_{ijnmp} = D_{jnmpi}I_{pnimj}$, it follows that if these two D's are automorphs, I_{pnimj} is also an automorph.

6. If $jm = jp = jn = ip = in = -ij$ *and* $mn = mp = 0$:
$C_{ijm,pn}I_{ijm}$ and $^*I_{-i,m}C_{mji,pn}I_{imj}$.

7. If $-ij = ip = in = im$: $^*E_{ij,pnm}$; also $E_{ij,x}$ where x is any permutation of pnm.

c) $a_{ij} > 0$, $a_{mn} > 0$ and other coefficients < 0.

1. If $ij + ip = 0$ see b)2.

2. If $ij + jn + mn = 0$: B_{ijnm} if $ij = mn$, $jm + mn = jp$, $in + nm = np$;
$B_{ijnm}I_{jn,im}$ if $in + nm = jp$, $ij + jm = np$ and $mp + nm = ij + ip$.

II. Second Type

Condition	Transformation
Always	$-I$
If $ip = jp$, $im = jm$, $in = jn$	$^*\pm I_{ij}$
If $ip = jp = ij$, $im = jm = pm$, $in = jn = pn$	$\pm I_{ijp}$
If $ip = jn$, $in = jp$, $im = jm$, $nm = pm$	$\pm I_{ij,pn}$
If $in = jn = pn = mn$, $ip = jm$, $ij = jp = pm = im$	$^*\pm I_{ijpm}$
If $ip = in = im = jp = jn = jm$, $pn = nm = pm$	$^*\pm I_{ij,pnm}$
If $ij = jp = pn = nm = mi$ and $ip = jn = pm = ni = mj$	$\pm I_{ijpnm}$

CORNELL UNIVERSITY.

ANNALS OF MATHEMATICS
Vol. 40, No. 1, January, 1939

PIECEMEAL UNIVALENCY OF ANALYTIC FUNCTIONS

BY M. S. ROBERTSON

(Received Feb. 23, 1938, revised May 2, 1938)

1. Introduction

The class (S_1) of functions

$$f(z) = z + \sum_2^\infty a_n z^n \tag{1.1}$$

regular and univalent for $|z| < 1$ has been studied by many authors. Several interesting properties for such functions have been found. To cite only three of them, we may mention the following inequalities:[1]

$$r(1+r)^{-2} \leqq |f(re^{i\theta})| \leqq r(1-r)^{-2}, \qquad r<1, \tag{1.2}$$

$$\frac{1}{2\pi}\int_0^{2\pi} |f(re^{i\vartheta})|\, d\theta < r(1-r)^{-1}, \qquad r<1, \tag{1.3}$$

$$\left|\frac{a_n}{n}\right| \leqq A < e, \tag{1.4}$$

where A is an absolute constant.

More recently several writers[2] have obtained results, analogous to those mentioned above, for the class (S_p) of functions

$$f(z) = \sum_{n=1}^\infty a_n z^n, \tag{1.5}$$

regular and multivalent of order p for $|z| < 1$. The corresponding inequalities, of necessity not as sharp as those above, are as follows:

$$|f(re^{i\theta})| < A(p)\mu_q(1-r)^{-2p}, \qquad r<1, \tag{1.6}$$

$$\frac{1}{2\pi}\int_0^{2\pi} |f(re^{i\theta})|\, d\theta < A(p)\mu_q(1-r)^{-2p+1}, \qquad r<1, \tag{1.7}$$

$$|an| < A(p)\mu_q n^{2p-1}, \tag{1.8}$$

where μ_q = maximum $\{|a_1|, |a_2|, \cdots, |a_q|\}$, q being the number of zeros of $f(z)$ located within the unit circle. $A(p)$ denotes a constant depending upon p but not upon $f(z)$.

---

[1] See L. Bieberbach, *Lehrbuch der Funktionentheorie*, Vol. 2, 1931, pp. 71–83.

[2] See M. Cartwright, *Some inequalities in the theory of functions*, Math. Annalen, Vol. 111, (1935), pp. 98–118. See also M. Biernacki, *Sur les fonctions multivalentes d'ordre p*, Paris Comptes Rendus, Vol. 203, (1936), pp. 449–451.

In this paper we investigate a subclass (S^α) of (S_p) of multivalent functions

$$f(z) = z^k + \sum_{k+1}^{\infty} a_n z^n, \qquad k \geqq 1, \tag{1.9}$$

regular for $|z| < 1$ which have the additional property of being univalent within pieces of the unit circle consisting of sectors. Let us state this property more explicitly in the following

DEFINITION. Let $f(z)$ be regular for $|z| < R$. If there is a positive real number $\alpha < 2$ such that $f(z)$ is univalent within every open sector of the circle $|z| = R$ of aperture $\alpha\pi$ and vertex at the center of the circle, then $f(z)$ is said to be *uniformly sectorally univalent of order* α with respect to the circle $|z| = R$.

Hereafter, the abbreviation u.s.u. (α) will be used for the phrase "uniformly sectorally univalent of order α."

We now suppose that the class (S^α) mentioned above consists of the functions $f(z)$ given by (1.9) which are u.s.u (α) with respect to the unit circle.

A very simple example of a function of class (S^α) which is not univalent in the unit circle as a whole, but which is u.s.u. (α) is z^k, $k > 1$, $\alpha = 2/k$. An example which is not univalent in the unit circle as a whole and which has $k = 1$ is the function

$$f(z) = z + z^2 + \frac{z^3}{3}.$$

This function has its radius of univalency equal to $\frac{\sqrt{3}}{2} < 1$. However, it is u.s.u. of order

$$\alpha = \frac{2}{\pi} \arccos\left(\frac{3 + \sqrt{4\rho^2 - 3}}{4\rho}\right)$$

with respect to the circle $|z| = \rho$ for $\frac{\sqrt{3}}{2} \leqq \rho < 1$. It is multivalent of order 2 within the unit circle. For this function $\alpha \to 0$ with $1 - \rho$.

We may note here that if $f(z)$ be u.s.u. (α) where $\alpha > 1$, and if also $f'(0) \neq 0$, then $f(z)$ is univalent throughout the whole circle in the ordinary sense. For if it were not, there would be two points z_1 and z_2 within the unit circle whose arguments differ by an amount not greater than π, and for which

$$f(z_1) - f(z_2) = 0.$$

This, however, would contradict the assumption that $f(z)$ is u.s.u. (α) where $\alpha > 1$.

The results analogous to (1.6), (1.7), and (1.8) obtained in this paper for the subclass (S^α) of the class (S_p) of multivalent functions are surprisingly much sharper in spite of the fact that the order of multivalency p may be very large. The rate of growth expressed by these inequalities for values of r near one is of the same order as that for the class (S_1) of univalent functions. Thus they

more nearly resemble the inequalities (1.2), (1.3) and (1.4) than (1.6), (1.7) and (1.8). Indeed, for the class (S^α) we have

$$|f(re^{i\theta})| < A(\alpha)r^k(1-r)^{-2}, \tag{1.10}$$

$$\frac{1}{2\pi}\int_0^{2\pi} |f(re^{i\theta})|\,d\theta < [2/\alpha]A(\alpha)r^k(1-r)^{-1}, \tag{1.11}$$

$$\left|\frac{a_n}{n}\right| < e[2/\alpha]A(\alpha), \qquad n > k, \tag{1.12}$$

where $A(\alpha)$ is a constant depending only upon α and not upon $f(z)$. We have used the same $A(\alpha)$ in (1.10), (1.11), and (1.12). The symbol $[x]$ denotes here the largest integer contained in the positive real number x.

Before proceeding with the proof of these statements let us observe that

$$f'(z) \neq 0, \qquad 0 < |z| < 1, \tag{1.13}$$

on account of the univalency property within every sector. $f'(z)$ may possibly vanish at the origin, however. Furthermore, we have

$$f(z) \neq 0, \qquad 0 < |z| < 1. \tag{1.14}$$

For if we suppose the contrary to be true, then $f(z_0)$ vanishes for some point z_0 within the circle and different from the origin. We make use of the fact that $f(z)$ takes on every value sufficiently near zero in a neighbourhood of z_0. Thus if we let Γ be the sector of aperture $\alpha\pi$ whose interior bisector passes through z_0 we can find a sufficiently small neighbourhood of the origin so that if z lies anywhere within that part of this neighbourhood which is common to the interior of Γ then $f(z)$ will take on a value there which it also takes on in a sufficiently small neighbourhood of z_0 within Γ. This being the case, $f(z)$ could not be univalent within Γ which is a contradiction. Hence, $f(z)$ can have no zeros within the unit circle other than at the origin.

Again, we should notice that $f(z)$ is at most multivalent of some finite order p within the unit circle whenever it is u.s.u. (α). This order p cannot exceed $[2/\alpha]$. To establish the truth of this statement let us observe that if $f(z)$ assumes the same value at two points z_1 and z_2 then the amplitudes of z_1 and z_2 must differ by an amount not less than $\alpha\pi$. For if not, z_1 and z_2 would lie interior to some sector of aperture $\alpha\pi$ within which $f(z)$ must be univalent. Hence, since there are not more than $[2/\alpha]$ points within the unit circle whose amplitudes each differ from one another by at least $\alpha\pi$, the order of multivalency cannot exceed $[2/\alpha]$.

2. The mapping of a circle onto a sector

Before going ahead with the proof of the statements of the preceding section of this paper, we need first to obtain the mapping function which will map in a conformal one-to-one fashion the interior of the unit circle $|u| = 1$ in the u-plane

into the interior of the sector Γ of aperture $\alpha\pi$ in the z-plane whose boundary consists of the two radii of the unit circle which are given by the equations

$$\arg z = \pm \frac{\alpha\pi}{2}$$

and of that arc of the unit circle which crosses the positive real axis.

Let $\phi(u)$ be defined by the equations

$$w(u) = \frac{(1+u^2)^{\frac{1}{2}} - 1}{u} = \frac{u}{2} + \cdots, \tag{2.1}$$

$$\begin{aligned}\phi(u) &= (\sqrt{2}-1)^{\alpha}\cdot\left[\frac{1+(\sqrt{2}+1)w(u)}{1-(\sqrt{2}-1)w(u)}\right]^{\alpha} \\ &= (\sqrt{2}-1)^{\alpha}[1 + 2\sqrt{2}\cdot\alpha w + \cdots],\end{aligned} \tag{2.2}$$

where $0 < \alpha \leqq 2$ and $(\sqrt{2}-1)^{\alpha}$ is a positive real number. The function $\phi(u)$ transforms $|u| = 1$ onto the required sector Γ so that the origin in the u-plane goes into the point $(\sqrt{2}-1)^{\alpha}$ of the positive real axis in the ϕ-plane. In order that $u = 0$ shall go into an arbitrary real point $a > 0$ within Γ select the real number x of absolute value less than one for which $\phi(x) = a$. Form the function

$$z = \phi\left(\frac{u+x}{1+xu}\right), \qquad |x| < 1. \tag{2.3}$$

By (2.3) the interior of $|u| = 1$ is mapped in a conformal one-to-one fashion into the interior of the sector Γ in the z-plane so that $u = 0$ corresponds to $z = a$. By taking the function inverse to $\phi(u)$ we find that the equation $\phi(x) = a$ has the unique solution

$$x = -\left(\frac{1-2b-b^2}{1+2b-b^2}\right) \tag{2.4}$$

where $b = a^{1/\alpha}$ is a positive real number.

3. The fundamental inequality for functions of class (S^{α})

With the notation of the preceding section let $f(z)$ be any function regular and univalent inside the open sector Γ of the z-plane. Since $f(z)$ is regular about the point $z = a$ we have

$$f(z) = \sum_{n=0}^{\infty} \frac{f^{(n)}(a)}{n!}(z-a)^n, \qquad f'(a) \neq 0, \tag{3.1}$$

$$z - a = \phi\left(\frac{u+x}{1+xu}\right) - \phi(u) = \sum_{1}^{\infty} \alpha_n u^n, \qquad \alpha_1 \neq 0, \tag{3.2}$$

$$f(z) - f(a) = \alpha_1 f'(a)u + \left\{\alpha_2 f'(a) + \frac{\alpha_1^2}{2} f''(a)\right\}u^2 + \cdots, \tag{3.3}$$

$$F(u) \equiv \frac{f(z) - f(a)}{\alpha_1 f'(a)} = u + A_2 u^2 + \cdots, \tag{3.4}$$

where

$$A_2 = \frac{\alpha_2}{\alpha_1} + \frac{\alpha_1}{2}\cdot\frac{f''(a)}{f'(a)}. \tag{3.5}$$

Since $f(z)$ is regular and univalent in Γ and since the function (3.2) is also univalent for $|u| < 1$ it follows that $F(u)$ is also univalent for $|u| < 1$. Since $F'(0) = 1$ we have as a consequence of a well known inequality of Bieberbach[3]

$$\left|\frac{\alpha_2}{\alpha_1} + \frac{\alpha_1}{2}\cdot\frac{f''(a)}{f'(a)}\right| = |A_2| \leqq 2. \tag{3.6}$$

Simple calculations give

$$\alpha_1 = \left(\frac{dz}{du}\right)_{u=0} = \phi'(x)(1-x^2),$$

$$\alpha_2 = \frac{1}{2}\left(\frac{d^2z}{du^2}\right)_{u=0} = \frac{(1-x^2)^2}{2}\cdot\phi''(x) - x(1-x^2)\phi'(x).$$

Hence (3.6) reduces to the form

$$\left|\frac{(1-x^2)}{2}\cdot\frac{\phi''(x)}{\phi'(x)} - x + \frac{(1-x^2)\phi'(x)}{2}\cdot\frac{f''(a)}{f'(a)}\right| \leqq 2. \tag{3.7}$$

Further calculations yield

$$\phi(x) = (\sqrt{2}-1)^\alpha\left\{\frac{1+(\sqrt{2}+1)w(x)}{1-(\sqrt{2}-1)w(x)}\right\}^\alpha,$$

$$w(x) = \frac{(1+x^2)^{\frac{1}{2}}-1}{x}, \qquad w'(x) = (1+(1+x^2)^{\frac{1}{2}})^{-1}\cdot(1+x^2)^{-\frac{1}{2}},$$

$$\phi'(x) = \frac{2\sqrt{2}(\sqrt{2}-1)^\alpha\alpha x^2\{x+(\sqrt{2}+1)((1+x^2)^{\frac{1}{2}}-1)\}^{\alpha-1}}{(1+x^2)^{\frac{1}{2}}(1+(1+x^2)^{\frac{1}{2}})\{x-(\sqrt{2}-1)((1+x^2)^{\frac{1}{2}}-1)\}^{\alpha+1}},$$

$$\frac{\phi''(x)}{\phi'(x)} = \{\alpha\sqrt{2}(1+x^2)^{\frac{1}{2}} - 2x^2 - x - 1\}(1+x)^{-1}(1+x^2)^{-1},$$

$$(1+x^2)^{\frac{1}{2}} = \frac{\sqrt{2}(b^2+1)}{1+2b-b^2}, \qquad 1-x^2 = \frac{8b(1-b^2)}{(1+2b-b^2)^2},$$

where $b = a^{1/\alpha}$.

Using (2.4) and the formulae above we finally have

$$\frac{1}{2}(1-x^2)\phi'(x) = \alpha\cdot a\left(\frac{1-b^2}{1+b^2}\right),$$

$$\frac{1}{2}(1-x^2)\frac{\phi''(x)}{\phi'(x)} - x = \alpha\left(\frac{1-b^2}{1+b^2}\right) - \frac{4b^2}{(1+b^2)^2}.$$

[3] See L. Bieberbach, loc. cit.

We may now write (3.7) as follows:

$$\left|\frac{af''(a)}{f'(a)} + 1 - \frac{4b^2}{\alpha(1-b^4)}\right| \leqq \frac{2}{\alpha}\left(\frac{1+b^2}{1-b^2}\right). \tag{3.8}$$

If we now make the additional assumption that $f(z)$ be regular and univalent within *every* sector of aperture $\alpha\pi$ of the unit circle, and a fortiori if $f(z)$ is a member of class (S^α) we then have the following fundamental inequality for all z for which $0 < |z| < 1$:

$$\left|\frac{zf''(z)}{f'(z)} + 1 - \frac{4|z|^{2/\alpha}}{\alpha(1-|z|^{4/\alpha})}\right| \leqq \frac{2}{\alpha}\left(\frac{1+|z|^{2/\alpha}}{1-|z|^{2/\alpha}}\right). \tag{3.9}$$

This follows from (3.8) on taking the function $f(ze^{i\theta})$ in place of $f(z)$ and evaluating it at $z = a > 0$.

If $f(z)$ belongs to (S^α) then (3.9) holds also for $0 \leqq |z| < 1$ since the left hand side then has a meaning on taking the limit as z approaches zero. Thus for $z = 0$ (3.9) gives

$$\alpha \leqq \frac{2}{k}, \qquad k \leqq [2/\alpha]. \tag{3.10}$$

$$\alpha \leqq \frac{2}{p} \leqq \frac{2}{k}, \qquad k \leqq p \leqq [2/\alpha]. \tag{3.11}$$

4. Proof of (1.10)

From (3.9) we have for $|z| = r < 1$

$$\Re\frac{zf''(z)}{f'(z)} + 1 - \frac{4r^{2/\alpha}}{\alpha(1-r^{2/\alpha})} \leqq \frac{2}{\alpha}\left(\frac{1+r^{2/\alpha}}{1-r^{2/\alpha}}\right),$$

$$\frac{r\partial}{\partial r}\log\{r^{1-2/\alpha}\cdot|f'(z)|\} = 1 - \frac{2}{\alpha} + \Re\frac{zf''(z)}{f'(z)}, \tag{4.1}$$

$$\frac{\partial}{\partial r}\log\{r^{1-2/\alpha}\cdot|f'(re^{i\theta})|\} \leqq \frac{4}{\alpha}\frac{2r^{2/\alpha-1}+r^{4/\alpha-1}}{1-r^{4/\alpha}}.$$

The inequality (4.1) is well known[4] for the special case $\alpha = 2$, that is for the class (S_1) of univalent functions. The equality sign in (3.9) and (4.1) is attained by the function $z^k(1-z^k)^{-2}$ where $k = 2/\alpha$. This function belongs to the class (S^α) if k is a positive integer.

Integrating (4.1) with respect to r from r_0 to r where $0 < r_0 < r < 1$ keeping θ constant, we have

$$\int_{r_0}^{r}\frac{\partial}{\partial r}\log\{r^{1-2/\alpha}\cdot|f'(re^{i\theta})|\}\,dr \leqq \frac{4}{\alpha}\int_0^r\frac{2r^{2/\alpha-1}+r^{4/\alpha-1}}{1-r^{4/\alpha}}\,dr,$$

[4] See L. Bieberbach, loc. cit.

whence for $0 < r_0 \leqq r < 1$

$$\left|\frac{f'(re^{i\theta})}{f'(r_0e^{i\theta})}\right| \leqq \left(\frac{r}{r_0}\right)^{2/\alpha-1} \cdot \frac{1+r^{2/\alpha}}{(1-r^{2/\alpha})^2}. \tag{4.2}$$

The special case $\alpha = 2/k$ is of interest for then

$$\lim_{r_0 \to 0} r_0^{1-2/\alpha} |f'(r_0e^{i\theta})| = k,$$

and (4.2) reduces to

$$|f'(re^{i\theta})| \leqq \frac{k(r^{k-1}+r^{2k-1})}{(1-r^k)^2}, \qquad r<1. \tag{4.3}$$

However, as $\alpha \leqq 2/k$ by (3.10) in general we may not permit r_0 to approach zero. To get around this difficulty let us proceeed as follows. Assuming that $f(z)$ belongs to (S^α) and is given by the power series (1.9) we know that $f(z)$ is multivalent of order p where $p \leqq [2/\alpha]$ as was pointed out in section one of this paper. Since also $f(z) \neq 0$ for $0 < |z| < 1$ we may therefore use a special case of the inequality (1.8) due to M. Biernacki[5] which will state that there exists a constant $A(p)$ independent of $f(z)$ so that for all n

$$|a_n| \leqq A(p)n^{2p-1} \tag{4.4}$$

Using (4.4) and (3.11) we can write for $0 \leqq r \leqq \frac{1}{2}$

$$\begin{aligned} |f'(re^{i\theta})| &\leqq kr^{k-1} + \sum_{k+1}^{\infty} n|a_n|r^{n-1} \leqq kr^{k-1} \cdot A(p) \sum_{n=k}^{\infty} \frac{n^{2p}}{2^{n-k}} \\ &\leqq kr^{k-1}A(p) \sum_{n=0}^{\infty} \frac{(n+p)^{2p}}{2^n} = kr^{k-1}\mu(p) \end{aligned} \tag{4.5}$$

where $\mu(p)$ is a constant depending upon p but not upon $f(z)$.

Let us now take $r_0 = \frac{1}{2}$ in the inequality (4.2). Because of (4.5) and (3.11) the inequality (4.2) may be written for $0 \leqq r < 1$ in the form

$$|f'(re^{i\theta})| < k2^{2/\alpha-k}\mu(p)\frac{r^{k-1}+r^{2k-1}}{(1-r^k)^2}, \qquad r<1. \tag{4.6}$$

On integrating (4.6) with respect to r from 0 to r keeping θ constant we also have

$$|f(re^{i\theta})| < \frac{A(\alpha)r^k}{(1-r)^2}, \qquad r<1, \tag{4.7}$$

where $A(\alpha) < 2^{2/\alpha-k}\mu(p)$. (4.7) is the desired inequality (1.10).

For the special case $\alpha = 2/k$ (4.7) can be obtained in a more exact form for we may integrate (4.3) and obtain

$$|f(re^{i\theta})| \leqq r^k(1-r^k)^{-2}. \tag{4.8}$$

This inequality is attained by the function $z^k(1-z^k)^{-2}$ which belongs to $(S^{2/k})$.

[5] See M. Biernacki, loc. cit.

In case $k = p$ we are able to obtain an estimate for $A(\alpha)$:

$$A(\alpha) < \frac{e^{8\pi p+2}}{\alpha^{2p}}, \qquad p \leqq [2/\alpha]. \tag{4.9}$$

In proving (4.9) we may assume $\alpha \neq 2/p$ and $p > 1$. By (2.3) $z = 0$ corresponds to a point $u = e^{i\nu}$. Since $F(u)$ of (3.4) is univalent for $|u| < 1$ and $F(u)$ is regular at the point $e^{i\nu}$ we have[6]

$$\left|\frac{f(a)}{\alpha_1 f'(a)}\right| = |F(e^{i\nu})| \geqq \frac{1}{4}. \tag{4.10}$$

Replacing α_1 by its value in terms of a and α and then changing the notation from a to z we have

$$\left|\frac{zf'(z)}{f(z)}\right| \leqq \frac{2}{\alpha}\left(\frac{1+r^{2/\alpha}}{1-r^{2/\alpha}}\right), \qquad |z| = r < 1. \tag{4.11}$$

In particular we have

$$r\frac{\partial}{\partial r}\log|f(re^{i\theta})| = \Re\frac{zf'(z)}{f(z)} \leqq \frac{2}{\alpha}\left(\frac{1+r^{2/\alpha}}{1-r^{2/\alpha}}\right). \tag{4.12}$$

On integrating from r_0 to r we have for $r_0 \leqq r < 1$

$$|f(re^{i\theta})| \leqq |f(r_0e^{i\theta})|\,\frac{(1-r_0^{2/\alpha})^2}{r_0^{2/\alpha}}\cdot\frac{r^{2/\alpha}}{(1-r^{2/\alpha})^2}. \tag{4.13}$$

Since however, $k = p$, $a_p = 1$, we have[7]

$$|f(r_0e^{i\theta})| \leqq e^{8\pi p}r_0^p(1-r_0)^{-2p}. \tag{4.14}$$

Hence (4.13) becomes for $0 \leqq r < 1$

$$|f(re^{i\theta})| \leqq \frac{e^{8\pi p}}{r_0^{2/\alpha-p}(1-r_0)^{2p}}\,r^p(1-r^p)^{-2}. \tag{4.15}$$

Let $r_0 = 1 - 2\,(2/\alpha - p + 2)^{-1}$. Then

$$\begin{aligned}|f(re^{i\theta})| &\leqq e^{8\pi p}\left[1+\frac{2}{2/\alpha-p}\right]^{2/\alpha-p}\cdot\left[\frac{1-(p/2-1)\alpha}{\alpha}\right]^{2p}\cdot r^k(1-r^p)^{-2}\\ &< \frac{e^{8\pi p+2}}{\alpha^{2p}}\cdot r^p(1-r^p)^{-2},\end{aligned} \tag{4.16}$$

$$\overline{\lim_{r\to 1}}\,(1-r)^2|f(re^{i\theta})| \leqq \frac{e^{8\pi p+2}}{p^2\alpha^{2p}}. \tag{4.17}$$

---

[6] See L. Bierberbach, loc. cit.

[7] See M. Cartwright, loc. cit.

5. The proof of (1.11) and (1.12)

We shall prove first the following

LEMMA. *If $f(z)$ be a member of class (S^α) and given by* (1.9) *then the function $F(z)$ defined as*

$$F(z) = \sqrt{f(z^2)} = z^k + \sum_{k+1}^{\infty} b_n z^n \tag{5.1}$$

is regular and u.s.u. (β) where $\beta \geqq \alpha/2$ with respect to the unit circle.

PROOF. Since $f(z)$ does not vanish for $0 < |z| < 1$ the function $\sqrt{f(z^2)}$ is regular for $|z| < 1$. Let z_1 and z_2 be any two distinct points within the unit circle whose amplitudes differ by less than $\alpha\pi/2$. Then the amplitudes of z_1^2 and z_2^2 differ by less than $\alpha\pi$. Since $f(z)$ is a member of (S^α) we have

$$F(z_1) - F(z_2) = \sqrt{f(z_1^2)} - \sqrt{f(z_2^2)} \neq 0.$$

Consequently $F(z)$ is u.s.u. (β) where $\beta \geqq \alpha/2$ with respect to the unit circle.

We should note here that $F(z)$, as defined by (5.1) is multivalent of order not exceeding $[2/\alpha]$. This follows since $F(z)$ has the same order of multivalency[8] as has $f(z)$.

If $F(z)$ is defined by (5.1) with the aid of (4.7) we obtain immediately the inequality:

$$|F(re^{i\theta})| < \frac{r^k\sqrt{A(\alpha)}}{(1-r^2)}, \qquad r < 1. \tag{5.2}$$

We now make use of J. E. Littlewood's method for the proof of (1.3), to obtain (1.11) and (1.12), taking care for the multivalency of $F(z)$, and using our new inequality (5.2). This gives

$$\begin{aligned} \frac{1}{2\pi}\int_0^{2\pi} |f(re^{i\theta})|\,d\theta &= \frac{1}{2\pi}\int_0^{2\pi} |F(\rho e^{i\theta})|^2\,d\theta, && \rho^2 = r, \\ &< [2/\alpha]\cdot\frac{2}{\pi}\int_0^{\rho} \pi \max_{|z|=\rho} |F(z)|^2 \frac{d\rho}{\rho} \\ &\leqq [2/\alpha]A(\alpha)\int_0^{\rho} \frac{2\rho^{2k-1}}{(1-\rho^2)^2}\,d\rho \\ &\leqq [2/\alpha]A(\alpha)r^k(1-r)^{-1}. \end{aligned} \tag{5.3}$$

Since also

$$|a_n|\,r^n \leqq \frac{1}{2\pi}\int_0^{2\pi} |f(re^{i\theta})|\,d\theta,$$

on taking $r = (1 - 1/n)$ and using (5.3) we obtain

$$\left|\frac{a_n}{n}\right| < e\,[2/\alpha]\,A(\alpha). \tag{5.4}$$

RUTGERS UNIVERSITY,
NEW BRUNSWICK, N. J.

[8] See M. S. Robertson, *Multivalent functions of order p*, Bull. Am. Math. Soc., Vol. 44, No. 4, (1938), pp. 282–285.

ANNALS OF MATHEMATICS
Vol. 40, No. 1, January, 1939

TWO-DIMENSIONAL METRIC SPACES WITH PRESCRIBED GEODESICS

BY HERBERT BUSEMANN

(Received May 9, 1938)

1. If S is a plane SL space, i.e. a two-dimensional complete metric space with exactly one geodesic (= shortest line = SL) of infinite length passing through two given points (see the precise definition in No. 6), then S can be mapped topologically onto the Euclidean Plane π[1]). Under this mapping the geodesics of the SL space will be transformed into curves α with the following properties:

(*) They are homeomorphic to Euclidean straight lines. Either of the branches issuing from a point on any one curve extends to infinity. Exactly one curve α passes through any two distinct points x and y. The segment $\bar{\alpha}_{xy}$ determined by two points x and y on α depends continuously on x and y.

The present paper is concerned with a kind of converse of the preceding assertion. Namely, it will be proved that:

(I) Given a system A of curves α in the Euclidean plane π, which have the properties (*), a distance-function xy, preserving the topology of π, can be found under which π is an SL space with the curves of A as shortest lines.[2]

The statement (I) can be used to find *a variational problem with given extremals.* For by restricting the nature of the system A in steps one can reduce gradually to the usual conditions.[3] However, a thorough discussion of the manner in which the smoothness of the variational problem constructed depends on the system A, is possible only after the local properties of variational integrands have been investigated more closely. This investigation will be given elsewhere. Here I shall restrict myself to giving a set of conditions covering the cases usually considered. From these conditions it will be clear at once that the integrand resulting from our construction is smooth and that the given curves minimize the corresponding integral. It seems to me that the result thus obtained has the advantage of showing immediately that one has a solution in the large, whereas the usual method requires a special investigation in each case.

2. Let π be the Euclidean plane with the Cartesian coördinates (x, y). The Euclidean distance $\sqrt{[(x_2 - x_1)^2 + (y_2 - y_1)^2]}$ between two points $p(x_1, y_1)$

[1] See [2]. The numbers [n] refer to the references at the end of this note.

[2] By means of Theorem (I) we may construct a plane SL space in which pairs of shortest lines occur such that one line of a pair is asymptotically parallel to the other without the second being asymptotically parallel to the first. This answers a question which I had reaised in [3], p. 235. See No. 4 of the present paper.

[3] See for instance [1], p. 30.

and $q(x_2, y_2)$ will be designated by $e(p, q)$. By $U(\nu, \epsilon)$, where ν is any set in π, we mean the set of all points x at a distance less than ϵ from ν. We now state the exact assumptions regarding our system A of curves α:

1) *Each curve α of A can be represented in the form*

$$x = x(t), \qquad y = y(t), \quad -\infty < t < \infty,$$

where $x(t)$ and $y(t)$ are continuous, and such that if $p(t) = (x(t), y(t))$, $p(t_1) \neq p(t_2)$ for $t_1 \neq t_2$,

$$e(p(t), p(o)) \to \infty \text{ for } |t| \to \infty,$$

2) *Through any two distinct points x, y of π there passes exactly one curve α_{xy} of A.*

3) *Designating by $\bar{\alpha}_{xy}$ the set of points on α_{xy} between x and y, x and y included, when $x \neq y$, and setting $\alpha_{xx} = x$ otherwise, we require finally: if $x_\nu \to x$ and $y_\nu \to y$, then for a given $\epsilon > 0$ an $N(\epsilon)$ can be found such that for $\nu > N(\epsilon)$*

$$\bar{\alpha}_{x_\nu y_\nu} \subset U(\bar{\alpha}_{xy}, \epsilon) \quad \text{and} \quad \bar{\alpha}_{xy} \subset U(\bar{\alpha}_{x_\nu y_\nu}, \epsilon).$$

Since $\bar{\alpha}_{xy}$ is closed one sees that, if $z_\nu \subset U(\alpha_{xy}, \epsilon_\nu)$ and $\epsilon_\nu \to 0$, then each limit-point z of $\{z_\nu\}$ is in α_{xy}. For there exists on α_{xy} a point t_ν with $e(z_\nu, t_\nu) < \epsilon_\nu$. If $z_{n_\nu} \to z$ then $e(z, t_{n_\nu}) < e(z, z_{n_\nu}) + e(z_{n_\nu}, t_{n_\nu}) \to 0$. This enables us to show:

If $x_\nu \to x$, $y_\nu \to y$, $x \neq y$ and $z_\nu \subset \alpha_{x_\nu y_\nu}$, then each limit-point z of z_ν is on α_{xy}, and each point s of α_{xy} is an accumulation-point of points $s_\nu \subset \alpha_{x_\nu y_\nu}$. Hereafter we shall say that $\alpha_{x_\nu y_\nu}$ tends to α_{xy}.

The proof is simply this. Let $z_{\nu_n} \to z$. For $z_{\nu_n} \subset \bar{\alpha}_{xy}$ the assertion follows from 3) and the remark just made. Therefore let $\{z_{\nu_m}\}$ be a subsequence of $\{z_{\nu_n}\}$ with $(z_{\nu_m}, x_{\nu_m}, y_{\nu_m})$[4] and assume that $z \notin \alpha_{xy}$. We consider α_{yz}. Since $z \notin \alpha_{xy}$ the curve α_{yz} is different from α_{xy} and therefore x is not on α_{yz}. But from 3) it follows that $\bar{\alpha}_{y_{\nu_m} z_{\nu_m}} \subset U(\alpha_{yz}, \epsilon_m)$ with $\epsilon_m \to 0$. Hence $x = \lim x_{\nu_m} \subset \alpha_{yz}$. That each point of α_{xy} is an accumulation point of points s_ν on $\alpha_{x_\nu y_\nu}$ follows easily from this consideration and 3).

A consequence thereof is the statement: Given any two different points x, y and two positive numbers ϵ and m, a δ can be found such that $\alpha_{rs} \cdot U(x, m) \subset U(\alpha_{xy}, \epsilon)$ for $e(r, x) < \delta$ and $e(s, y) < \delta$.

Now let the points $r_1, r_2, \cdots$ be everywhere dense in π and $(r_1^1, r_1^2), (r_2^1, r_2^2), \cdots$ any ordering of the pairs of different r_ν. Then, according to the last statement the curves $\alpha_{r_\nu^1 r_\nu^1}$ are dense in A in the following sense: *given $\epsilon > 0$, $m > 0$ and two points x, y, a ν_0 can be found such that*

$$\alpha_{xy} \cdot U(x, m) \subset U(\alpha_{r_{\nu_0}^1 r_{\nu_0}^2}, \epsilon).$$

We shall now show that the curves of A have as many of the elementary properties of Euclidean straight lines as we shall need later on.

We call a point set in π *α-convex* if it contains $\bar{\alpha}_{xy}$ when it contains x and y.

---

[4] (a, b, c) means that b is on α_{ac} and between a and c.

Each curve α in A decomposes π into exactly two open regions ρ and ρ', the sides of α. $\rho + \alpha$ and $\rho + \alpha'$ are α-convex. For if x and y are on α, then $\bar{\alpha}_{xy} \subset \alpha$. If $x \subset \alpha$, $y \subset \rho$, then $\bar{\alpha}_{xy}$ cannot contain a point z of ρ', since $\bar{\alpha}_{xy}$ would contain a point $s \neq x$ of α, and α_{xy} would have two intersections with α. If, finally, $x \subset \rho$, $y \subset \rho$ and yet $\bar{\alpha}_{xy}$ contained a point z of ρ', then $\bar{\alpha}_{xz}$ and $\bar{\alpha}_{zy}$ would each contain a point of α.

The point p decomposes $\alpha_{pq} (p \neq q)$ into two open rays; the one containing q we designate by $\vec{\alpha}_{pq}$. From the last proof it follows:

If $p \subset \alpha$ and $q \subset \rho$, then $\vec{\alpha}_{pq} \subset \rho$.

We wish to show that, if (r, p, q), then $\vec{\alpha}_{pr}$ lies in ρ'.

Assume $r \subset \rho$; then $\vec{\alpha}_{pr} \subset \rho$. $\vec{\alpha}_{pq}$, $\vec{\alpha}_{pr}$ decompose ρ, and the side ρ' of α is completely contained in one of the sides, ρ_1' say, of α_{pq}, whereas the other side is contained in ρ. Since $\rho + \alpha$ and $\rho_1' + \alpha_{pq}$ are convex, their intersection β (which is not empty because α is in it) is convex. β is therefore connected, but $\beta - p$ is not connected. It consists of two disjoint connected parts, q and r belonging to different parts. Let $r' \subset \alpha$ belong to the same part as r; then r' and q belong to different parts. Hence $\bar{\alpha}_{qr}$, must pass through p. This means that $\alpha_{qr'}$, and α have the two different points p and r' in common.

We can express our result as follows:

If α and α_1 intersect at p, then the two rays into which $\alpha(\alpha_1)$ decomposes $\alpha_1(\alpha)$ are on different sides of $\alpha(\alpha_1)$.

We see furthermore that the two sides ρ and ρ' of a curve α in A are α-convex. If α_1 has the sides ρ_1 and ρ_1' and $\alpha \cdot \alpha_1 = p$, then $\alpha + \alpha_1$ decomposes π into four α-convex regions, namely $\rho \cdot \rho_1$, $\rho_1 \cdot \rho_1'$, $\rho' \cdot \rho_1'$, $\rho_1' \cdot \rho$. The point p and any two rays α-rays α^+ and α_1^+ issuing from a point p but not belonging to the same α-curve, decompose the plane π into two open regions, one of which is α-convex and the other not. The convex region will be called the angle with the sides α^+ and α_1^+. If $l \subset \alpha^+$ and $l_1 \subset \alpha_1^+$ we designate this angle by (α^+, α_1^+) or $\measuredangle lpl_1$. $(\alpha^+, \alpha_1^+) + \alpha^+ + \alpha_1^+ + p$ is convex and closed; it contains $\bar{\alpha}_{ll_1}$; and the interior points of this α-segment are contained in (α^+, α_1^+).

Let ρ be that side of α_{ll_1} which contains p and let ρ' be the other side. The intersection $\beta = \rho \cdot (\alpha^+, \alpha_1^+)$ is not empty, for if (lkp) and (l_1k_1p) hold, then the interior points of $\bar{\alpha}_{kk_1}$ are in ρ as well as in (α^+, α_1^+). Hence β is open, α-convex, and connected. On the other hand, since $\delta = \bar{\alpha}_{ll_1} + \bar{\alpha}_{l_1p} + \bar{\alpha}_{pl}$, is a simple closed Jordan curve, it decomposes π into exactly two domains. Take s and s_1 such that (sll_1) and (ll_1s_1). Then

$$\pi - \delta - \beta = \rho' + \vec{\alpha}_{ls} + \vec{\alpha}_{l_1s_1} + \measuredangle slp + \measuredangle s_1l_1p.$$

If $x \subset \rho'$, then $\bar{\alpha}_{xs} \subset \rho' + \vec{\alpha}_{ls}$ and $\bar{\alpha}_{xs_1} \subset \rho' + \vec{\alpha}_{l_1s_1}$. If $y \subset \measuredangle slp$, then $\bar{\alpha}_{ys} \subset \measuredangle slp + \bar{\alpha}_{ls}$ and if $z \subset \measuredangle s_1l_1p$ then $\bar{\alpha}_{zs_1} \subset \measuredangle s_1l_1p + \vec{\alpha}_{l_1s_1}$. Therefore $\pi - \delta - \beta$ is connected; hence β must be the interior of δ. From the preceding discussion one easily concludes that the curves of A have all those properties of the Euclidean straight lines, which are usually derived from the *axioms of order and connexion* in the foundations of geometry. If, in particular p is in the interior

of the α-triangle abc then each ray issuing from p cuts $\bar{\alpha}_{ab} + \bar{\alpha}_{bc} + \bar{\alpha}_{ca}$ exactly once. Therefore, the rays issuing from p, considered as elements of a set, can be mapped topologically on the circumference of a circle. After having fixed a positive sense of rotation in π we may speak of revolving the ray $\vec{\alpha}_{px}$ continuously around p in the positive sense.

3. Now let $p \notin \alpha_{ab}$; take q such that (q, p, b); and let x traverse $\vec{\alpha}_{ba} + b$ monotonically from b to infinity. Then $\vec{\alpha}_{px}$ turns around p monotonically but never passes $\vec{\alpha}_{pq}$. Hence α_{px} tends to a limit-curve α in A which we call an *asymptote* to α_{ab} through p; in particular, we call the limit-ray of $\vec{\alpha}_{px}$ on α the *asymptote* to $\vec{\alpha}_{ba}$ through p. There are two possibly coinciding asymptotes α and α_1 through p to α_{ab} . If they coincide, α is the only curve of A through p which does not intersect α_{ab} ; in this case we call α a *parallel* to α_{ab} through p. If α is different from α_1 and if $\alpha^+(\alpha_1^+)$ denotes that ray of $\alpha(\alpha_1)$ issuing from p which is on the same side of $\alpha_1(\alpha)$ as α_{ab} , the rays issuing from p and lying in (α^+, α_1^+) intersect α_{ab} and the other rays issuing from p do not.

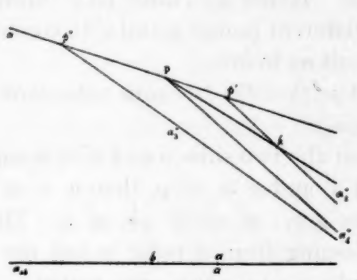


FIG. 1

If $p' \subset \alpha^+$ then α is also an asymptote to α_{ab} through p'. We note first that two non-intersecting curves of A decompose π into three parts, one of which itself decomposes the plane. Each subset of this last set will be said to be between the two curves of A. We continue as follows: if our assertion were not true, the ray α_2^+ through p', which is the asymptote to $\vec{\alpha}_{ab}$ through p', would be between α and α_{ab} . If $k \subset \alpha_2^+$, then $\vec{\alpha}_{pk}$ would also be between α and α_{ab} and would not intersect α_{ab} . Then α^+ could not be the asymptote to $\bar{\alpha}_{ab}$ through p.

If (p'', p, p'), α is also an asymptote to $\vec{\alpha}_{ab}$ through p''. If $\alpha_3^+ \supset \alpha^+$ were the ray asymptotic to $\vec{\alpha}_{ab}$ through p'' this ray would again be between α and α_{ab} . The ray α_4^+ asymptotic to α_3^+ through p would be different from α^+, since α and α_3 intersect at p''. Because (p'', p, p'), α_4^+ is between α and α_{ab} ; so that α^+ could not be the asymptote to $\vec{\alpha}_{ab}$ through p. We see that:

If α is an asymptote to $\vec{\alpha}_{ab}$ through one of its points, then it is an asymptote to $\vec{\alpha}_{ab}$ through each of its points. We therefore call α an asymptote to $\vec{\alpha}_{ab}$ without mentioning a point on it. If we call α_{ab} an asymptote to itself, we can say that through each point of the plane there passes exactly one asymptote to $\vec{\alpha}_{ab}$.

These asymptotes are continuous in the following sense:

If $p_r \to p$ then the asymptote α_r to $\vec{\alpha}_{ab}$ through p_r tends to the asymptote α to $\vec{\alpha}_{ab}$ through p.

We have to show that if $q_{r_n} \subset \alpha_{r_n}$ and $q_{r_n} \to q$, then $q \subset \alpha$. Assume $q \not\subset \alpha$ and let α' be the asymptote to $\vec{\alpha}_{ab}$ through q. Since $q \neq p$, α_{pq} crosses α and α' at q and p respectively. But because $\alpha_{p r_n q r_n}$ tends to α_{pq} it would have to cross α and α' too, whereas different asymptotes to $\vec{\alpha}_{ab}$ have no common points.

4. We see that many of the properties of asymptotes in Euclidean and hyberbolic geometries hold for any system A of curves α satisfying conditions 1), 2), and 3). There is, however, an important property of the asymptotes which holds in the elementary cases but fails to hold in the general case, namely, *the concept of an asymptote is no longer symmetric*, i.e., if a ray α_1^+ is asymptotic to the ray α_2^+, then α_2^+ is not necessarily an asymptote to α_1^+.

In order to construct an example which will prove this assertion we consider, in an (x, y)-plane, the branch $x < 0$ of the hyperbola $xy = 1$. We designate it by δ'. δ' has $x = 0$ and $y = 0$ as asymptotes in the usual sense. Moreover, there exists exactly one tangent of δ' parallel to a given straight line (SL) $y = ax$, $a < 0$, and there exists no tangent parallel to an SL $y = ax$ with $a \geqq 0$, or to $x = 0$. If we make a secant T_x traverse monotonically all the secants parallel to a fixed tangent T to δ', then, starting with T, the lengths of the chords cut out by T_x increase monotonically in the strict sense from 0 to ∞, also, there are no other chords of δ' parallel to T. Designating by D the family of curves δ derived from δ' by translations $x' = x + c$, $y' = y + d$ and designating by pq the Euclidean SL through p and q, we can say:

Corresponding to each pair of points p, q, $p + q$, such that pq has an equation of the form

I. $$y = ax + b, \qquad a < 0,$$

there exists exactly one curve δ in D passing through p and q.

Let E be the family of all SL, whose equations can be written either in form

II $$y = ax + b, \qquad a \geqq 0 \quad \text{or} \quad \text{II}' \quad x = e.$$

Then $A = D + E$ is a system of curves satisfying the conditions 1), 2), and 3), but for which *the property of being an asymptote is, in general, not symmetric.* For, take the point $(-1, -1)$. The asymptote to the negative y-axis through this point is our curve δ'. The asymptote to the ray $y < -1$ of the SL $x = -1$ through $(0, 0)$ is the y-axis. Hence this last ray and the negative y-axis form a pair of rays of which the latter is asymptotic to the first, without the first ray being asymptotic to the second.

One should notice that to each line II there exists a parallel through every point of the plane, and to a line $x = e$ there exists a parallel through each point (x, y) with $x > e$, but there exist two different asymptotes through each point with $x < e$.

The system A is transformed into itself by all translations of the plane, and, as we shall see later, there exists an SL space for which the curves of A are the geodesics or shortest lines. It is, however, remarkable that there is no SL space, such that, in addition, the distances in it remain unchanged under the translations of π. This is an immediate consequence of theorem (4.1) in [3].

5. We return now to the general considerations of no. 3. Let O be the origin of a system of Cartesian coördinates x, y and K_{2n} the circle $x^2 + y^2 = 4n^2$. Choose the points $x_1, x_2, \cdots x_m$ in this order on K_{2n} so densely that $\bar{\alpha}_{x_\nu x_{\nu+1}}$ ($\nu = 1, \cdots, m$; $x_{m+1} = x_1$) is contained in $U(x_\nu, n/4)$. Since K_{2n} separates O from infinity, $\rho = \sum_{\nu=1}^{m} \bar{\alpha}_{x_\nu x_{\nu+1}}$ does too. The proof follows the usual lines by showing that the indices of O and infinity with respect to K_{2n} are the same as with respect to ρ.[5] Since $\rho \cdot U(0, n) = 0$, ρ separates all points of $U(0, n)$ from infinity.

We shall now show that ρ has a bounded α-convex closure; by this we mean the smallest α-convex set β containing ρ. Such a set exists, namely, the intersection of all α-convex sets containing ρ. Furthermore, β is obviously also the α-convex closure of the set $x_1 + \cdots + x_m$. We fix any ray α^+ issuing from O and draw the asymptotes to α^+ through $x_1 \cdots, x_m$. Since of three asymptotes to α^+ one is always between the other two, there is one, say α_1 among these m asymptotes such that the others either coincide with α_1 or are all on the same side ρ_1 of α_1. We now fix a sense on α_1. With respect to this orientation let x_{i_1} be the first and x_{i_2} the last of the points x_i on α_1. Let α_1^+ be the ray of those points of α_1 which follow x_{i_2} and let α_1^- be the ray of the points of α_1 preceding x_{i_2}. Define a positive sense of rotation in π such that α_1^+ is turned around x_{i_2} through ρ_1 into the position α_1^- by a rotation in the positive sense. Now revolve α_1^+ positively until it meets for the first time a point x_ν and call x_{i_3} the last point x_ν following x_{i_2} on the ray thus obtained. Then revolve the ray $\vec{\alpha}_{x_{i_2}x_{i_3}} - \bar{\alpha}_{x_{i_2}x_{i_3}}$ positively around x_{i_3} until it hits the next point x_ν; the last point x_ν on this ray will be x_{i_4}; and so forth. This process must end after a finite number of steps, say l, and leads back to x_{i_1}, such that $x_{i_l} = x_{i_1}$. Therefore $\delta = \sum_{\nu=1}^{l-1} \bar{\alpha}_{x_{i_\nu} x_{i_{\nu+1}}}$ is a polygon which bounds an α-convex region. This region together with δ contains ρ and $U(0, n)$. We see that:

There exists a sequence of polygons $\delta_1, \delta_2, \cdots$ bounding α-convex regions, and such that $\delta_{\nu+1}$ contains δ_ν and $U(0, \nu + 1)$ in its interior.

We now make the following construction: We map δ_1 topologically onto the unit circle $r = 1$ of an (r, φ) plane so that points on the same α-curve through O are mapped on diametrically opposite points of the circle. Let α_φ^+ be the ray issuing from O and passing through the point of δ_1 corresponding to φ. Consider now δ_ν and $\delta_{\nu+1}$ and draw all α-rays $\alpha_1^+, \cdots, \alpha_{s+1}^+ = \alpha_1^+$ issuing from 0 and passing through a vertex of δ_ν or $\delta_{\nu+1}$. Let α_i^+ intersect δ_ν in y_i^0 and $\delta_{\nu+1}$ in y_i^1.

[5] Comp. [4] p. 84.

We map $\alpha_{y_i^0 y_i^1}$ topologically onto $0 \leqq t \leqq 1$ in such a way that y_i^0 corresponds to 0 and y_i^1 to 1. The point corresponding to t will be denoted by y_i^t. We draw all the α-segments $\bar{\alpha}_{y_i^t y_{i+1}^t}$, $0 < t < 1$, $i = 1, \cdots, s$. The polygon $\delta_{r+t} = \sum_{i=1}^{s} \bar{\alpha}_{y_i^t y_{i+1}^t}$, $0 \leqq t \leqq 1$, is star-shaped with respect to O. For O and δ_1 we make an analogous construction, connecting O to the vertices of δ_1, and define thus α-polygons δ_t for $0 < t < 1$ also. Each point of π, except O, is on exactly one polygon δ_r, $0 < r < \infty$, and on exactly one ray α_φ^+. We map the point on δ_r and on α_φ^+ on the point (r, φ) of the (r, φ) plane. If we complete this mapping by associating O with the point $r = 0$ of the (r, φ) plane, we get a homeomorphism under which the α-curves through O are mapped on the straight lines through $r = 0$. Mapping the (r, φ) plane by $(r, \varphi) \to \left(\frac{r}{r+1}, \varphi\right)$ onto the interior β of the unit circle, we have finally a representation of our system A with the following properties:

The α-curves through O are the diameters of β; for each circle $r = c < 1$ or $e(0, x) = c$ there exists a closed α-polygon bounding an α-convex region containing the circle; either end of any α-curve tends towards a definite point of the unit-circle.

That the last property holds, is a consequence of the fact that if by traversing the curve α in one sense, one did not approach a definite point, then one would approach the points of a whole arc of the unit circle; and a radius of β, ending at an interior point of this arc, would have infinitely many points in common with α.

From now on we hold to this representation of A. Besides polar coördinates we introduce Cartesian coördinates in β. The segment $-1 < x < 1$ of $y = 0$ is a curve α_0 of A; call α_0^+ the ray $x < 0$ of α_0. We consider the family of asymptotes to α_0^+. If α and α' are any two such asymptotes in $y > 0$, one of them, α say, separates the other from α_0. There exists a uniquely determined Euclidean circle around O which has points in common with α, but whose interior is completely on one side of α. Let $t(\alpha)$ be the radius of this circle. Of course $t(\alpha)$ is the distance of α from O. We have $t(\alpha) < t(\alpha')$. In the same manner there corresponds a radius t to each asymptote α to α_0^+ in $y < 0$; however, in this case we put $t(\alpha) = -t$. We complete this definition by putting $t(\alpha_0) = 0$. The important properties of this association are:

(a) *to different asymptotes α, α' there correspond different values $t(\alpha)$ and $t(\alpha')$.*

(b) *If α' is between α and α'', then either*

$$t(\alpha) > t(\alpha') > t(\alpha'') \quad \text{or} \quad t(\alpha) < t(\alpha') < t(\alpha'').$$

It is clear that $t(\alpha_\nu) \to t(\alpha)$ for $\alpha_\nu \to \alpha$. We must know furthermore that there belongs an asymptote to each value t between -1 and 1. One sees first that, if there belongs an asymptote to each of the values t_ν, and $t_\nu \to t_0$, $|t_0| < 1$, then an asymptote corresponds to t_0 also. For let p_ν be on α_ν and on the circle of radius $t_\nu = t(\alpha_\nu)$ around O and let $\{p_{\nu_n}\}$ be a converging subsequence of $\{p_\nu\}$.

Its limit p_0 is on the limit α_0 of $\{\alpha_{\nu_n}\}$ and on the circle with radius t_0. α_0 cannot contain an interior point of this circle, since such a point cannot be an accumulation-point of points $q_\nu \subset \alpha_\nu$; therefore, $t_0 = t(\alpha_0)$. Thus the set of values t corresponding to the asymptotes is closed relatively to $-1 < t < 1$. Its complement is open or empty, but the first possibility can easily be excluded with the help of the fact that $t(\alpha_\nu) \to t(\alpha)$ for $\alpha_\nu \to \alpha$. We have therefore:

(c) *For each t, $-1 < t < 1$, there exists exactly one asymptote $\alpha(t)$ to α_0^+ and vice versa; for $t_\nu \to t$ one has $\alpha(t_\nu) \to \alpha(t)$ and for $\alpha_\nu \to \alpha$ one has $t(\alpha_\nu) \to t(\alpha)$.*

6. For this section it is not essential that the curves $\alpha(t)$ should be a family of asymptotes to the same ray. We shall only use the property that each curve α of A belongs to a system F of curves in A such that through a given point q there passes exactly one curve α_q of F with the property $\alpha_{q_n} \to \alpha_q$ for $q_n \to q$, and F satisfies the conditions (a), (b), (c). Using an artifice it is now easy to construct a metric which makes the curves of A the shortest lines of an SL space. We first give the exact definition of an SL space:

A complete metric space with the distance function $\rho(x, y)$ is an SL space if to each pair of points a, b there exists exactly one point c, the center of a and b, for which $\rho(a, c) + \rho(c, b) = \rho(a, b)$ and $\rho(a, c) = \rho(c, b)$, and, furthermore, there exist exactly two points d and d' such that b is the center of a and d and a is the center of b and d'.

For three different points a, b, c, we express the relation $\rho(a, b) + \rho(b, c) = \rho(a, b)$ by $[a, b, c]$. In this number we shall show then, that a metric preserving the topology of β can be found such the relations (a, b, c) and $[a, b, c]$ are equivalent for all triples of distinct points.

We start with a family F of curves α in A as described before. Through any point p there passes exactly one curve α_p of F; the number $t(\alpha_p)$ therefore is uniquely determined by p. We put

$$d(p, q) = |\, t(\alpha_p) - t(\alpha_q)\,| .$$

We have (1) $0 \leqq d(p, q) < 2$.

(2) $d(p, q) = 0$ if, and only if, $\alpha_p = \alpha_q$.

(3) $d(p, q) + d(q, r) \geqq d(p, r)$

(4) $d(p, q) + d(q, r) = d(p, r)$ either if α_q coincides with one of the lines α_p or α_r , or if α_q is between α_p and α_r .

(5) $d(p, q) + d(q, r) > d(p, r)$, if α_p , α_q , α_r are different and α_q is not between α_p and α_r , or if $\alpha_p = \alpha_r \neq \alpha_q$.

$d(p, q)$ depends on the family F. Each curve α belongs to such a family. Let now $\alpha_\nu = \alpha_{\nu_\nu^1 \nu_\nu^2}$ be a countable set of α-curves dense in A, as constructed in

no. 2. Let α_ν belong to the family F_ν. To each F_ν we associate a function $d_\nu(p, q)$ satisfying conditions (1), · · · (5). We are going to prove that

$$r(p, q) = \sum_{\nu=1}^{\infty} \frac{d_\nu(p, q)}{2^{\nu+2}}$$

defines a metric space for which *the relations* (a, b, c) *and* $[a, b, c]$ *are equivalent.* $r(p,q)$ is finite on account of (1). The relations $r(p, q) = 0$ for $p = q$ and $r(p, q) + r(q, r) \geqq r(p, r)$ are immediate consequences of (2) and (3). $r(p, q) > 0$ for $p \neq q$ can be shown thus: Call $\alpha_{p,\nu}$ the curve of F_ν through p. If $p \neq q$ then a curve α in A exists separating p from q. Hence also some curve α_j of $\{\alpha_\nu\}$ will separate p from q, and $\alpha_{j,p}$ and $\alpha_{j,q}$ will also be separated by α_j ; (2) shows that $d_j(p, q) > 0$. In order to prove that (p, q, r) if, and only if, $[p, q, r]$, assume first (p, q, r). If $\alpha_{\nu,p} = \alpha_{\nu,r}$, then $d_\nu(p, q) = d_\nu(q, r) = 0$. If $\alpha_{\nu,p} \neq \alpha_{\nu,r}$, the domain bounded by these two curves is α-convex and q is in that domain. Hence $\alpha_{\nu,q}$ is between $\alpha_{\nu,p}$ and $\alpha_{\nu,r}$, and we have, on account of (4), that $d_\nu(p, q) + d_\nu(q, r) = d_\nu(p, r)$. By summing we see that $r(p, q) + r(q, r) = r(p, r)$ or $[p, q, r]$. Now let $q \notin \bar{\alpha}_{pr}$. Then a curve in A exists which separates q from $\bar{\alpha}_{pr}$. Therefore, a curve α_j can be found in $\{\alpha_\nu\}$ separating q from $\bar{\alpha}_{pr}$. Hence q cannot be between $\alpha_{j,p}$ and $\alpha_{j,r}$ nor can it coincide with $\alpha_{j,p}$ or $\alpha_{j,r}$. It follows then from (3) and (5) that $r(p, q) + r(q, r) > r(p, r)$.

It will be clear that any two points p and q possess a unique center if we can prove that $r(p, x)$ changes continuously as x traverses $\bar{\alpha}_{pq}$ from p to q. It is easy to see that $r(p, x) \to 0$, if $x \to p$. For one has $\alpha_{\nu,x} \to \alpha_{\nu,p}$ for $x \to p$; hence, on account of property (c) of F_ν, $d_\nu(p, x) \to 0$ for $x \to p$. $d_\nu(p', q') < 2$ for any p' and q', and the definition of $r(p', q')$ show then that $r(p, x) \to 0$ for $x \to p$. That $x \to p$ for $r(x, p) \to 0$ is contained in the considerations of the preceding paragraph; so that the metric $r(p', q')$ is topologically equivalent with the Euclidean metric. From this follows, in particular, that $r(p, x)$ changes continuously as x traverses $\bar{\alpha}_{pq}$. *We thus have obtained a metric* $r(\ )$ *topologically equivalent to* $e(\ )$, *for which the curves of A are the shortest lines, but* $r(\ )$ *does not define an SL space, since* $r(\ ) < 2$, *and in an SL space the shortest lines must have infinite lengths.*

7. In order to meet this last requirement, we have to make a construction which is a little more complicated than the one used before, but based on the same principle. Take again the previous representation of A in the interior β of the unit circle. The α-rays issuing from 0 are the Euclidean radii of β. Let α^+_φ, $0 \leqq \varphi < 2\pi$, be these radii. Take a definite one, α^+_ψ say, and draw the two asymptotes α_1, α_2 to α^+_ψ which touch the circle $e(0, x) = 1 - 1/n$, $n > 1$. Let $a_i \subset \alpha_i$ decompose α_i into the two rays α^+_i and α^-_i $(i = 1, 2)$, where the notation is such that α^+_1 and α^+_2 are on the same side of $\alpha_{a_1 a_2}$ as are the points of α^+_ψ sufficiently remote from O. Designate by $p^\pm_i$ the point on the unit-circle which one approaches as one traverses $\alpha^\pm_i$. Then p^-_i is different from $(1, \psi)$. For otherwise one would

have $p_i^+ = p_i^- = (1, \psi)$ and there would be radii of β intersecting α_i twice. Now take the asymptotes δ_1, δ_2 to α_ψ^+ through points b_1, b_2 with (a_2, a_1, b_1) and (a_1, a_2, b_2), respectively. Let $q_i^\pm$ be the endpoints of δ_i, $q_i^+ \neq q_i^-$; one of the open arcs of the unit circle founded by q_i^+ and q_i^- does not contain $(1,\psi)$ in the interior. Call δ_ψ the open set consisting of these two arcs $q_i^+ q_i^-$. To each point $(1, \varphi)$ of the unit-circle there belongs such an arc S_φ; we can find a finite number of points $(1, \varphi_1), \cdots, (1, \varphi_k)$ such that the corresponding arcs $S_{\varphi_1}, \cdots, S_{\varphi_k}$ cover the unit-circle.

We now define a distance as follows: Let $(1, \psi)$ be one of these points with the notations as before. To each circle $e(0, x) = t < 1$ there belong two uniquely determined asymptotes to α_ψ^+ touching this circle. Making the same association

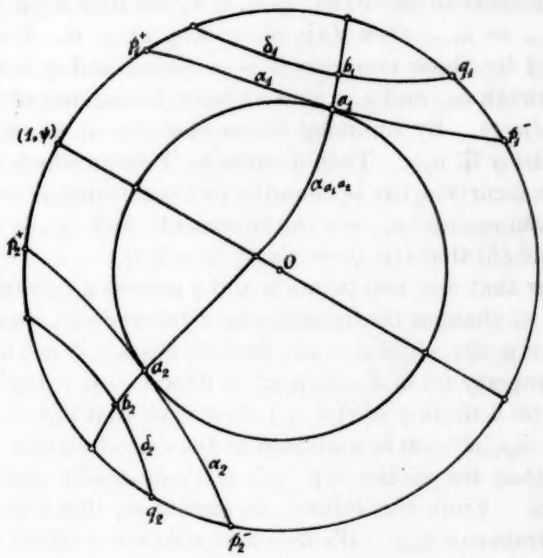


FIG. 2

as in no. 5 we get a function $t(\alpha)$, where the values t between -1 and 1 are in one-to-one continuous correspondence with asymptotes to α_ψ^+. Let $t(\delta_i) = t_i$. We have $t_1 > 1 - 1/n > -1 + 1/n > t_2$ (or this relation with t_1 and t_2 interchanged). We form a strictly monotonic function $f(t)$ defined for $-1 < t < 1$ with $f(1 - 1/n) = 1/(k \cdot 2^n)$, $f(-1 + 1/n) = -1/(k \cdot 2^n)$, $f(t_1) = 1, f(t_2) = -1$. If p is on the asymptote to α_φ^+ corresponding to t' and q on the one corresponding to t'', we put

$$d'(p, q) = |f(t') - f(t'')|.$$

We remark that for all pairs p, q between α_1 and α_2, and therefore in particular for the pairs p, q in $U(0, 1 - 1/n)$, we have the inequality

$$d'(p, q) < \frac{2}{k \cdot 2^n},$$

and that each ray α^+ issuing from a point of $U(0, 1 - 1/n)$ and ending at a point of S_ψ has length at least $1 - 1/(1.2^n)$.

The function d has been constructed with respect to a definite $S_{\varphi i}$. We construct a similar function for every one of these sets $S_{\varphi i}$. We get the functions $d_1'(\ ), \cdots, d_k'(\ )$. We then put

$$d^n(p, q) = \sum_{i=1}^{k} d_i'(p, q).$$

We have $d^n(p, q) \subset 2/2^n$ for p, q in $U(0, 1 - 1/n)$, and the length of any ray issuing from a point of $U(0, 1 - 1/n)$ is at least $1 - 1/(1.2^n) > 1/2$. We form

$$\rho'(p, q) = \sum_{n=2}^{\infty} d^n(p, q).$$

This sum converges for every pair of points p, q. For if n_0 is sufficiently large we have $p + q \subset U(0, 1 - 1/n)$ for $n > n_0$. Therefore

$$\rho'(p, q) < \sum_{n=2}^{n_0} d^n(p, q) + \sum_{n=n_0+1}^{\infty} d^n(p, q) < \sum_{n=2}^{n_0} d^n(p, q) + 1/2^{n_0-1}.$$

Let α^+ be any ray leading from the point r to the point r' of the unit circle and $r \subset U(0, 1 - 1/n_0)$. If $r_\nu \to r'$ on α^+ and $N > 0$ is given, we can choose $\nu_0(N)$ so large, that for $\nu > \nu_0(N)$,

$$d^{n+n_0+1}(r, r_\nu) > \tfrac{1}{2}, \qquad n = 1, 2, \cdots, 2N.$$

Then $\rho'(r, r_\nu) > N$ for $\nu > \nu_0(N)$. As in number 6, we see that the $\rho(p, q)$ defined by

$$\rho(p, q) = \rho'(p, q) + r(p, q),$$

where $r(\ )$ is the metric defined in that number, *converts the interior β of the unit circle into a metric space with the curves of A as shortest lines. $\rho(p, q)$ converts β into an SL space since every α-ray has infinite length.*

8. The method developed here can be applied *to find a variational problem with given extremals.* For instance, let A be a system of curves satisfying conditions 1), 2), and 3) of no. 2; assume, furthermore, that each curve α in A has a continuous tangent and that through a given line element there passes exactly one curve of A. Each curve α belongs to a one-parameter-family $F = \alpha(t)$ of curves in A as described in no. 6, such that through any point r there passes exactly one curve $\alpha(t_r)$ of F. We assume that t_r is a differentiable function of the arc-length on $\bar{\alpha}$ as r traverses a curve $\bar{\alpha}$ in A. For every $\bar{\alpha}$ we associate this derivative of t_r with the line element consisting of r and the direction φ of $\bar{\alpha}$ at r. Thus we get a function $f(r, \varphi)$ defined for all line elements in β. We finally assume that $f(r, \varphi)$ is a continuous function of r and φ. Evidently, if p and q are any two points, one has

$$d(p, q) = |t(\alpha_p) - t(\alpha_q)| = \int_{\bar{\alpha}_{pq}} f(r, \varphi)\, ds,$$

r traversing $\bar{\alpha}_{pq}$, φ being the direction of α_{pq} at r, and ds being considered positive.

As in No. 6 we now take a set $\{\alpha_\nu\}$ of curves of A and dense in A. Each α_ν belongs to a family F_ν with the properties indicated. The functions $f(r, \varphi)$ and $d(p, q)$ belonging to F_ν we designate by $f_\nu(r, \varphi)$ and $d_\nu(p, q)$ respectively. Let $|f_\nu|$ be less than some suitable R_ν for $r < 1 - 1/\nu$ and all φ. Then we form

$$r(p, q) = \sum_{\nu=1}^{\infty} \frac{d_\nu(p, q)}{(1 + R_\nu)2^{\nu+2}}.$$

As before, $r(p, q)$ converts β into a metric space with the curves of A as unique shortest lines. Setting

$$F(r, \varphi) = \sum_{\nu=1}^{\infty} \frac{|f_\nu(r, \varphi)|}{(1 + R_\nu)2^{\nu+2}},$$

one has

$$r(p, q) = \int_{\bar{\alpha}_{pq}} F(r, \varphi)\, ds,$$

and $F(r, \varphi)$ depends continuously on r and φ. It is easy to see that $\bar{\alpha}_{pq}$ minimizes $\int_C F(r, \varphi) \cdot ds$ among all piecewise continuously differentiable curves C connecting p to q. Actually $\bar{\alpha}_{pq}$ *minimizes this integral among all rectifiable curves between* p *and* q. A proof for this more general fact is contained in the results of K. Menger in [5].

INSTITUTE FOR ADVANCED STUDY.

REFERENCES

[1] BOLZA, O., *Lectures on the Calculus of Variations*, Chicago 1904.

[2] BUSEMANN, H., *Pasch'sches Axiom und Zweidimensionalität*, Math. Ann. vol. 107 (1932), pp. 324–328.

[3] BUSEMANN, H., *On Translations in General Plane Geometrics*, Amer. Jour. of Math., vol. LX (1938), pp. 227–256.

[4] v. KERÉKJÁRTÓ, *Vorlesungen über Topologie I*, Berlin 1923.

[5] MENGER, K., *Die metrische Methode in der Variationsrechnung*, Erg. eines math. Kolloquiums, Heft 8, Wien 1937.

ANNALS OF MATHEMATICS
Vol. 40, No. 1, January, 1939

ON UNITARY METRICS IN PROJECTIVE SPACE

BY HERMANN WEYL

(Received September 29, 1938)

For the purpose of investigating meromorphic curves, my son Joachim and I introduced in the k-dimensional projective space R with the homogeneous coördinates $x_0, \cdots, x_k$ a unitary metric[1] by defining the distance $\| \xi x \|$ of a plane ξ and a point x and the distance $[[xx']]$ of two points x, x' by means of the equations

$$(1_1) \qquad \| \xi x \|^2 = \Big| \sum_i \xi_i x_i \Big|^2 : \Big(\sum_i |\xi_i|^2 \sum_i |x_i|^2 \Big),$$

$$(1_2) \qquad [[xx']]^2 = \sum_{i<j} |x_i x_j' - x_j x_i'|^2 : \Big(\sum_i |x_i|^2 \cdot \sum_i |x_i'|^2 \Big).$$

The question arose to what extent these distances depend on the coördinate system. Let us therefore subject the x to an arbitrary non-singular linear transformation

$$(2) \qquad y_i = \sum_j h_{ij} x_j$$

while the plane coördinates ξ_i undergo the contragredient transformation

$$(3) \qquad \xi_i = \sum_j h_{ji} \eta_j .$$

By passing to the new coördinate system y the two distances $\| \xi x \|$, $[[xx']]$ take on a factor varying between fixed positive limits, independent of ξ and x, or of x and x' respectively. We shall here determine the *exact* limits of both distances and discuss in more detail this problem which seems to deserve some attention on its own merits.

I use the notation of matrix calculus, designating by x the column of the numbers $x_0, x_1, \cdots, x_k$ (vector) and by A^* the transposed matrix of A. Unitary geometry is based upon the "scalar product"

$$(x, x') = \bar{x}_0 x_0' + \cdots + \bar{x}_k x_k' = \bar{x}^* x'.$$

In particular

$$(x, x) = |x_0|^2 + \cdots + |x_k|^2.$$

---

[1] Annals of Math. *39* (1938), pp. 516-538 (quoted as M). I take this opportunity to correct a few misprints:
p. 517, line 14: read "Γ^2 and γ^2" instead of "Γ and γ".
p. 533, line 15: cancel the factor k in $\dfrac{k(k+1)}{k-l+1}$.
p. 538, line 6 (last line of the Wronskian): read "$x^{(k)}$" instead of "$x^{(k-1)}$."

The inequality

$$|(x, x')|^2 \leqq (x, x)\cdot(x', x') \tag{4}$$

proves that both our distances (1) lie between 0 and 1. The numerator in (1_2) may be written as the determinant

$$\begin{vmatrix} (x, x) & (x, x') \\ (x', x) & (x', x') \end{vmatrix}.$$

By the transformation (2), $y = Hx$, one gets

$$(y, y) = (Hx, Hx) = \bar{x}^* \bar{H}^* H x = (x, Cx),$$

where

$$C = \bar{H}^* H$$

is Hermitian:

$$\bar{C}^* = C \quad \text{or} \quad (x, Cx') = \overline{(x', Cx)}.$$

The vectors $e_0, \cdots, e_k$ form a unitary basis if

$$(e_i, e_j) = \delta_{ij} = \begin{cases} 1 & (i = j) \\ 0 & (i \neq j) \end{cases}.$$

One knows that with a given Hermitian transformation C one is able to determine a unitary basis e_i such that

$$Ce_i = \lambda_i e_i.$$

We arrange the real eigen-values λ_i which are the roots of the secular equation

$$\det(\lambda E - C) = 0 \qquad (E = \text{unit matrix})$$

in increasing order, $\lambda_0 \leqq \lambda_1 \leqq \cdots \leqq \lambda_k$. One then has for any vector

$$x = t_0 e_0 + \cdots + t_k e_k \qquad (t_i \text{ numbers})$$

the equations

$$\begin{aligned} (x, x) &= |t_0|^2 + \cdots + |t_k|^2, \\ (x, Cx) &= \lambda_0 |t_0|^2 + \cdots + \lambda_k |t_k|^2 \end{aligned} \tag{5}$$

(transformation onto principal axes). This shows at once that

$$\lambda_0 \cdot (x, x) \leqq (x, Cx) \leqq \lambda_k \cdot (x, x).$$

In our case $C = \bar{H}^* H$, the Hermitian form (x, Cx) is positive definite and therefore all eigen-values λ_i are positive. We set

$$\lambda_0 = \gamma^2, \qquad \lambda_k = \Gamma^2 \tag{6}$$

and then have

$$\gamma^2 \cdot (x, x) \leqq (y, y) \leqq \Gamma^2 \cdot (x, x). \tag{7}$$

γ^2 and Γ^2 are the minimum and maximum of (y, y) under the condition $(x, x) = 1$.

All this is very well known. In denoting by d_x, d_y the distances in the old and the new coördinate system, we have for the *point-plane-distance* (1_1):

$$\left(\frac{d_x}{d_y}\right)^2 = \frac{\sum |y_i|^2}{\sum |x_i|^2} \cdot \frac{\sum |\eta_i|^2}{\sum |\xi_i|^2}. \tag{8}$$

By (7) we have succeeded in determining the exact bounds of the first factor. In the same fashion we should be able to compute those limits for the second factor. However, it is superfluous to go through the same process again; our transformation to principal axes settles the question concerning both factors. Indeed, let us designate the column of the $\bar{\xi}_i$ by ξ and the column of the $\bar{\eta}_i$ by η so that

$$\xi = \bar{H}^* \eta.$$

The maximum and minimum of the quotient

$$\frac{\sum |\xi_i|^2}{\sum |\eta_i|^2} = \frac{(\xi, \xi)}{(\eta, \eta)} \tag{9}$$

are the greatest and least eigen-values of

$$D = H\bar{H}^*.$$

But these eigen-values are the same as for

$$C = \bar{H}^* H,$$

while the eigen-vectors of D are

$$f_i = He_i/\sqrt{\lambda_i}.$$

Indeed one realizes at once that

$$(f_i, f_j) = \delta_{ij}, \qquad Df_i = \lambda_i f_i.$$

Hence the coördinates τ_i as introduced by

$$\eta = \tau_0 f_0 + \cdots + \tau_k f_k$$

satisfy the relations

$$(\eta, \eta) = |\tau_0|^2 + \cdots + |\tau_k|^2,$$
$$(\xi, \xi) = \lambda_0 |\tau_0|^2 + \cdots + \lambda_k |\tau_k|^2,$$

and thus the exact bounds of (9) turn out the same, γ^2 and Γ^2, as for the first factor in (8). The final result is

$$\gamma/\Gamma \leqq d_x/d_y \leqq \Gamma/\gamma. \tag{10}$$

It should be observed that the transformations $x_i \to t_i$, $\xi_i \to \tau_i$ are by no means contragredient. Instead one has

$$\sum_i \xi_i x_i = \sum_i \eta_i y_i = \sum_i \sqrt{\lambda_i} \cdot \bar{\tau}_i t_i,$$

as follows readily from

$$\sum_i \eta_i y_i = (\eta, Hx) = \sum_{i,j} \bar{\tau}_i t_j (f_i, He_j) = \sum_i \sqrt{\lambda_i} \cdot \bar{\tau}_i t_i .$$

The investigation for the *point distance* $d_x = [[xx']]$ first follows the same track: After a suitable unitary transformation we obtain

$$(11_x) \qquad d_x^2 = \begin{vmatrix} (x,x) & (x,x') \\ (x',x) & (x',x') \end{vmatrix} : (x,x)(x',x'),$$

$$(11_y) \qquad d_y^2 = \begin{vmatrix} L(x,x) & L(x,x') \\ L(x',x) & L(x',x') \end{vmatrix} : L(x,x)L(x',x'),$$

where

$$L(x,x') = \sum_i \lambda_i \bar{x}_i x_i' .$$

This time we have been less pedantic by allowing the new variables t_i to be denoted again by x_i. The numerator in d_y^2 may be written as

$$\sum_{i<j} \lambda_i \lambda_j |z_{ij}|^2$$

with

$$(12) \qquad z_{ij} = x_i x_j' - x_j x_i' .$$

This sum differs from

$$\sum_{i<j} |z_{ij}|^2$$

by a factor between the bounds $\lambda_0\lambda_1$ and $\lambda_{k-1}\lambda_k$. Together with

$$\lambda_0 \cdot (x,x) \leqq L(x,x) \leqq \lambda_k \cdot (x,x)$$

and the same inequality for x' this results in the estimate

$$(13) \qquad \frac{\lambda_0^2}{\lambda_{k-1}\lambda_k} \leqq \left(\frac{d_x}{d_y}\right)^2 \leqq \frac{\lambda_k^2}{\lambda_0\lambda_1} .$$

On replacing $\lambda_{k-1}\lambda_k$ by the bigger λ_k^2 and $\lambda_0\lambda_1$ by the lesser λ_0^2 one finds that d_x/d_y certainly lies between $(\Gamma/\gamma)^2$ and $(\gamma/\Gamma)^2$ as maintained in the paper M. But this rough estimate is far from being exact. I am going to prove that the precise limits are Γ/γ and γ/Γ again.

In the one-dimensional case $k = 1$ this follows at once from (13) because the left- and right-hand sides of this inequality equal

$$\lambda_0/\lambda_1 \quad \text{and} \quad \lambda_1/\lambda_0$$

respectively. The lower limit is attained when x and x' coincide at the point 1:0, and the upper limit when $x = x' = (0:1)$.

On returning to arbitrary k, this result may be applied to any straight line

in R. For any two points x, x' on $\mathfrak{z}$ the distance quotient d_x/d_y varies between two reciprocal limits $\kappa = \kappa(\mathfrak{z}) \geqq 1$ and $1/\kappa$ each of which will be attained if x, x' coincide at a certain point on the straight line. I proceed to compute κ explicitly. When the straight line is thought of as determined by two points x, x', its coördinates z_{ij} will be given by (12). In (11_x), (11_y) we have to substitute for x, x' two arbitrary points of this straight line, viz.

$$\sigma_1 x + \sigma_2 x', \qquad \tau_1 x + \tau_2 x' \tag{14}$$

respectively. In denoting the matrix

$$\begin{Vmatrix} L(x,x) & L(x,x') \\ L(x',x) & L(x',x') \end{Vmatrix} \text{ by } \begin{Vmatrix} L_{11} & L_{12} \\ L_{21} & L_{22} \end{Vmatrix}$$

and the corresponding one with $\lambda_i = 1$ by

$$\begin{Vmatrix} l_{11} & l_{12} \\ l_{21} & l_{22} \end{Vmatrix}$$

the squared distance d_y^2 of the two points (14) will come out as

$$(L_{11}L_{22} - L_{12}L_{21}) \mid \sigma_1\tau_2 - \sigma_2\tau_1 \mid^2 / L(\sigma)\cdot L(\tau)$$

where $L(\sigma)$ is the Hermitian form

$$L_{11}\bar{\sigma}_1\sigma_1 + L_{12}\bar{\sigma}_1\sigma_2 + L_{21}\bar{\sigma}_2\sigma_1 + L_{22}\bar{\sigma}_2\sigma_2$$

of two variables σ_1, σ_2. Hence

$$\left(\frac{d_x}{d_y}\right)^2 = \frac{L(\sigma)\cdot L(\tau)}{l(\sigma)\cdot l(\tau)} \cdot \frac{\det(l)}{\det(L)}.$$

The maximum and minimum of the Hermitian form $L(\sigma)$ under the condition $l(\sigma) = 1$ are the two roots λ of the secular equation

$$\begin{vmatrix} \lambda l_{11} - L_{11}, & \lambda l_{12} - L_{12} \\ \lambda l_{21} - L_{21}, & \lambda l_{22} - L_{22} \end{vmatrix} = 0.$$

The limits of d_x/d_y on the straight line under consideration are therefore the two values

$$\kappa = \lambda \cdot \left(\frac{\det l}{\det L}\right)^{\frac{1}{2}}$$

which are the roots of the equation

$$\kappa^2 - \frac{l_{11}L_{22} + l_{22}L_{11} - l_{12}L_{21} - l_{21}L_{12}}{\sqrt{\det L}\cdot\sqrt{\det l}}\kappa + 1 = 0.$$

This confirms our former statement that the two bounds κ, $1/\kappa$ are reciprocal, and by computing the middle coefficient we get

$$(15) \qquad \kappa + \frac{1}{\kappa} = \sum_{i<j} (\lambda_i + \lambda_j) z_{ij} \bar{z}_{ij} / \sqrt{\sum_{i<j} |z_{ij}|^2} \cdot \sqrt{\sum_{i<j} \lambda_i \lambda_j |z_{ij}|^2}.$$

After this piece of elementary calculation our next step will be to ascertain the maximum of the quotient at the right side, with the straight line $\mathfrak{z} = (z_{ij})$ taking on all possible positions. For any two numbers $\kappa \geqq 1$, $\kappa_0 \geqq 1$ the inequality

$$\kappa \leqq \kappa_0 \quad \text{implies} \quad \kappa + \frac{1}{\kappa} \leqq \kappa_0 + \frac{1}{\kappa_0},$$

and vice versa. Instead of determining the maximum κ_0 of $\kappa(\mathfrak{z})$ and hence the minimum $1/\kappa_0$ of $1/\kappa(\mathfrak{z})$ we therefore proceed to calculate the maximum $\kappa_0 + \frac{1}{\kappa_0}$ of $\kappa + \frac{1}{\kappa}$. To this end we apply our fundamental inequality (4),

$$\Big| \sum_p \bar{u}_p v_p \Big|^2 \leqq \sum_p |u_p|^2 \cdot \sum_p |v_p|^2,$$

in the following way. Let p range over the pairs ij with $i < j$, and set

$$u_p = z_{ij}, \qquad v_p = (\lambda_i + \lambda_j) z_{ij}.$$

We thus get

$$\Big\{ \sum_{i<j} (\lambda_i + \lambda_j) z_{ij} \bar{z}_{ij} \Big\}^2 \leqq \sum_{i<j} |z_{ij}|^2 \cdot \sum_{i<j} (\lambda_i + \lambda_j)^2 |z_{ij}|^2.$$

The square of the quotient (15) is therefore

$$\leqq \sum_{i<j} (\lambda_i + \lambda_j)^2 |z_{ij}|^2 / \sum_{i<j} \lambda_i \lambda_j |z_{ij}|^2.$$

On putting

$$\lambda_i \lambda_j |z_{ij}|^2 = V_p, \qquad \rho_p = \frac{\lambda_i + \lambda_j}{(\lambda_i \lambda_j)^{\frac{1}{2}}} = \left(\frac{\lambda_j}{\lambda_i}\right)^{\frac{1}{2}} + \left(\frac{\lambda_i}{\lambda_j}\right)^{\frac{1}{2}},$$

we resort to the trivial inequality

$$\sum_p \rho_p^2 V_p \leqq \rho^2 \cdot \sum_p V_p$$

where ρ is the biggest of the $k(k+1)/2$ numbers ρ_p, and we thus find the quotient (15) to be

$$\leqq \max_{i<j} \left[\left(\frac{\lambda_j}{\lambda_i}\right)^{\frac{1}{2}} + \left(\frac{\lambda_i}{\lambda_j}\right)^{\frac{1}{2}} \right].$$

This upper bound for $\kappa + \frac{1}{\kappa}$ implies

$$\kappa(\mathfrak{z}) \leqq \max_{i<j} \sqrt{\lambda_j/\lambda_i} = \sqrt{\lambda_k/\lambda_0} = \Gamma/\gamma$$

as we maintained.[2]

Any l points

$$x, x', \cdots, x^{(l-1)} \qquad (1 \leqq l \leqq k)$$

determine an $(l-1)$-spread with the coördinates

$$x_{i_1 \cdots i_l} = \begin{vmatrix} x_{i_1}, \cdots, x_{i_l} \\ \cdots\cdots\cdots\cdots \\ x_{i_1}^{(l-1)}, \cdots, x_{i_l}^{(l-1)} \end{vmatrix}.$$

Any skew-symmetric numbers $x_{i_1 \cdots i_l}$ dependent on the coördinate system which transform like these determinants describe a $(l-1)$-complex $x(l)$. The $(l-1)$-complexes form a projective space R_l of k_l dimensions,

$$k_l + 1 = \binom{k+1}{l}.$$

Spreads are special complexes. In R_l one can introduce a unitary metric based on the form

$$|x(l)|^2 = \sum_{i_1<\cdots<i_l} |x_{i_1 \cdots i_l}|^2. \tag{16}$$

For special complexes the formula

$$|x(l)|^2 = \begin{vmatrix} (x,x), & (x,x'), & \cdots, & (x,x^{(l-1)}) \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ (x^{(l-1)},x), & (x^{(l-1)},x'), & \cdots, & (x^{(l-1)},x^{(l-1)}) \end{vmatrix} \tag{17}$$

holds. We compare (16) with the corresponding expression in a new coördinate system as introduced by (2). According to the method formerly employed we may assume without loss of generality that

$$(y,y) = \lambda_0 |x_0|^2 + \cdots + \lambda_k |x_k|^2.$$

We then obtain from the expression (17) for $y(l)$, first for special complexes and then in general,

$$|y(l)|^2 = \det_{\alpha,\beta=0,\cdots,l-1} \left\{ \sum_i \lambda_i \bar{x}_i^{(\alpha)} x_i^{(\beta)} \right\} = \sum_{i_1<\cdots<i_l} \lambda_{i_1} \cdots \lambda_{i_l} |x_{i_1 \cdots i_l}|^2.$$

[2] Instead of adopting our elementary method one could resort to the infinitesimal process of comparing locally at any point the lengths of an arbitrary line element in the two metrics.

Thus the values Γ_l , γ_l similar to Γ and γ in R_l are given by

$$\gamma_l^2 = \lambda_0\lambda_1 \cdots \lambda_{l-1}\,, \qquad \Gamma_l^2 = \lambda_{k-l+1} \cdots \lambda_{k-1}\lambda_k\,.$$

From this exact estimate one may derive, if one so desires, the simpler though less complete result

$$\gamma_l \geqq \gamma^l, \qquad \Gamma_l \leqq \Gamma^l.$$

INSTITUTE FOR ADVANCED STUDY
PRINCETON, N. J.

ANNALS OF MATHEMATICS
Vol. 40, No. 1, January, 1939

ON UNITARY REPRESENTATIONS OF THE INHOMOGENEOUS LORENTZ GROUP*

BY E. WIGNER

(Received December 22, 1937)

1. ORIGIN AND CHARACTERIZATION OF THE PROBLEM

It is perhaps the most fundamental principle of Quantum Mechanics that the system of states forms a *linear manifold*,[1] in which a unitary *scalar product* is defined.[2] The states are generally represented by wave functions[3] in such a way that φ and constant multiples of φ represent the same physical state. It is possible, therefore, to normalize the wave function, i.e., to multiply it by a constant factor such that its scalar product with itself becomes 1. Then, only a constant factor of modulus 1, the so-called phase, will be left undetermined in the wave function. The linear character of the wave function is called the superposition principle. The square of the modulus of the unitary scalar product (ψ, φ) of two normalized wave functions ψ and φ is called the transition probability from the state ψ into φ, or conversely. This is supposed to give the probability that an experiment performed on a system in the state φ, to see whether or not the state is ψ, gives the result that it is ψ. If there are two or more different experiments to decide this (e.g., essentially the same experiment,

---

* Parts of the present paper were presented at the Pittsburgh Symposium on Group Theory and Quantum Mechanics. Cf. Bull. Amer. Math. Soc., 41, p. 306, 1935.

[1] The possibility of a future non linear character of the quantum mechanics must be admitted, of course. An indication in this direction is given by the theory of the positron, as developed by P. A. M. Dirac (Proc. Camb. Phil. Soc. *30*, 150, 1934, cf. also W. Heisenberg, Zeits. f. Phys. *90*, 209, 1934; *92*, 623, 1934; W. Heisenberg and H. Euler, ibid. *98*, 714, 1936 and R. Serber, Phys. Rev. *48*, 49, 1935; *49*, 545, 1936) which does not use wave functions and is a non linear theory.

[2] Cf. P. A. M. Dirac, *The Principles of Quantum Mechanics*, Oxford 1935, Chapters I and II; J. v. Neumann, *Mathematische Grundlagen der Quantenmechanik*, Berlin 1932, pages 19–24.

[3] The wave functions represent throughout this paper states in the sense of the "Heisenberg picture," i.e. a single wave function represents the state for all past and future. On the other hand, the operator which refers to a measurement at a certain time t contains this t as a parameter. (Cf. e.g. Dirac, l.c. ref. 2, pages 115–123). One obtains the wave function $\varphi_s(t)$ of the Schrödinger picture from the wave function φ_H of the Heisenberg picture by $\varphi_s(t) = \exp(-iHt/\hbar)\varphi_H$. The operator of the Heisenberg picture is $Q(t) = \exp(iHt/\hbar)Q\exp(-iHt/\hbar)$, where Q is the operator in the Schrödinger picture which does not depend on time. Cf. also E. Schrödinger, Sitz. d. Kön. Preuss. Akad. p. 418, 1930.

The wave functions are complex quantities and the undetermined factors in them are complex also. Recently attempts have been made toward a theory with real wave functions. Cf. E. Majorana, Nuovo Cim. *14*, 171, 1937 and P. A. M. Dirac, in print.

performed at different times) they are all supposed to give the same result, i.e., the transition probability has an invariant physical sense.

The wave functions form a description of the physical state, not an invariant however, since the same state will be described in different coördinate systems by different wave functions. In order to put this into evidence, we shall affix an index to our wave functions, denoting the Lorentz frame of reference for which the wave function is given. Thus φ_l and $\varphi_{l'}$ represent the same state, but they are different functions. The first is the wave function of the state in the coördinate system l, the second in the coördinate system l'. If $\varphi_l = \psi_{l'}$ the state φ behaves in the coördinate system l exactly as ψ behaves in the coördinate system l'. If φ_l is given, all $\varphi_{l'}$ are determined up to a constant factor. Because of the invariance of the transition probability we have

$$|(\varphi_l, \psi_l)|^2 = |(\varphi_{l'}, \psi_{l'})|^2 \tag{1}$$

and it can be shown[4] that the aforementioned constants in the $\varphi_{l'}$ can be chosen in such a way that the $\varphi_{l'}$ are obtained from the φ_l by a linear unitary operation, depending, of course, on l and l'

$$\varphi_{l'} = D(l', l)\varphi_l . \tag{2}$$

The unitary operators D are determined by the physical content of the theory up to a constant factor again, which can depend on l and l'. Apart from this constant however, the operations $D(l', l)$ and $D(l_1', l_1)$ must be identical if l' arises from l by the same Lorentz transformation, by which l_1' arises from l_1. If this were not true, there would be a real difference between the frames of reference l and l_1. Thus the unitary operator $D(l', l) = D(L)$ is in every Lorentz invariant quantum mechanical theory (apart from the constant factor which has no physical significance) completely determined by the Lorentz transformation L which carries l into $l' = Ll$. One can write, instead of (2)

$$\varphi_{Ll} = D(L)\varphi_l . \tag{2a}$$

By going over from a first system of reference l to a second $l' = L_1 l$ and then to a third $l'' = L_2 L_1 l$ or directly to the third $l'' = (L_2 L_1) l$, one must obtain—apart from the above mentioned constant—the same set of wave functions. Hence from

$$\varphi_{l''} = D(l'', l')D(l', l)\varphi_l$$
$$\varphi_{l''} = D(l'', l)\varphi_l$$

it follows

$$D(l'', l')D(l', l) = \omega D(l'', l) \tag{3}$$

---

[4] E. Wigner, *Gruppentheorie und ihre Anwendungen auf die Quantenmechanik der Atomspektren*. Braunschweig 1931, pages 251–254.

or

$$D(L_2)D(L_1) = \omega D(L_2L_1), \tag{3a}$$

where ω is a number of modulus 1 and can depend on L_2 and L_1. Thus the $D(L)$ form, up to a factor, a representation of the inhomogeneous Lorentz group by linear, unitary operators.

We see thus[5] that there corresponds to every invariant quantum mechanical system of equations such a representation of the inhomogeneous Lorentz group. This representation, on the other hand, though not sufficient to replace the quantum mechanical equations entirely, can replace them to a large extent. If we knew, e.g., the operator K corresponding to the measurement of a physical quantity at the time $t = 0$, we could follow up the change of this quantity throughout time. In order to obtain its value for the time $t = t_1$, we could transform the original wave function φ_l by $D(l', l)$ to a coördinate system l' the time scale of which begins a time t_1 later. The measurement of the quantity in question in this coördinate system for the time 0 is given—as in the original one—by the operator K. This measurement is identical, however, with the measurement of the quantity at time t_1 in the original system. One can say that the representation can replace the equation of motion, it cannot replace, however, connections holding between operators at one instant of time.

It may be mentioned, finally, that these developments apply not only in quantum mechanics, but also to all linear theories, e.g., the Maxwell equations in empty space. The only difference is that there is no arbitrary factor in the description and the ω can be omitted in (3a) and one is led to real representations instead of representations up to a factor. On the other hand, the unitary character of the representation is not a consequence of the basic assumptions.

The increase in generality, obtained by the present calculus, as compared with the usual tensor theory, consists in that no assumptions regarding the field nature of the underlying equations are necessary. Thus more general equations, as far as they exist (e.g., in which the coördinate is quantized, etc.) are also included in the present treatment. It must be realized, however, that some assumptions concerning the continuity of space have been made by assuming Lorentz frames of reference in the classical sense. We should like to mention, on the other hand, that the previous remarks concerning the time-parameter in the observables, have only an explanatory character, and we do not make assumptions of the kind that measurements can be performed instantaneously.

We shall endeavor, in the ensuing sections, to determine all the continuous[6] unitary representations up to a factor of the inhomogeneous Lorentz group, i.e., all continuous systems of linear, unitary operators satisfying (3a).

---

[5] E. Wigner, l.c. Chapter XX.

[6] The exact definition of the continuous character of a representation up to a factor will be given in Section 5A. The definition of the inhomogeneous Lorentz group is contained in Section 4A.

2. Comparison With Previous Treatments and Some Immediate Simplifications

A. Previous treatments

The representations of the Lorentz group have been investigated repeatedly. The first investigation is due to Majorana,[7] who in fact found all representations of the class to be dealt with in the present work excepting two sets of representations. Dirac[8] and Proca[8] gave more elegant derivations of Majorana's results and brought them into a form which can be handled more easily. Klein's work[9] does not endeavor to derive irreducible representations and seems to be in a less close connection with the present work.

The difference between the present paper and that of Majorana and Dirac lies—apart from the finding of new representations—mainly in its greater mathematical rigor. Majorana and Dirac freely use the notion of infinitesimal operators and a set of functions to all members of which every infinitesimal operator can be applied. This procedure cannot be mathematically justified at present, and no such assumption will be used in the present paper. Also the conditions of reducibility and irreducibility could be, in general, somewhat more complicated than assumed by Majorana and Dirac. Finally, the previous treatments assume from the outset that the space and time coördinates will be continuous variables of the wave function in the usual way. This will not be done, of course, in the present work.

B. Some immediate simplifications

Two representations are *physically equivalent* if there is a one to one correspondence between the states of both which is 1. invariant under Lorentz transformations and 2. of such a character that the transition probabilities between corresponding states are the same.

It follows from the second condition[5] that there either exists a unitary operator S by which the wave functions $\Phi^{(2)}$ of the second representation can be obtained from the corresponding wave functions $\Phi^{(1)}$ of the first representation

$$\Phi^{(2)} = S\Phi^{(1)} \tag{4}$$

or that this is true for the conjugate imaginary of $\Phi^{(2)}$. Although, in the latter case, the two representations are still equivalent physically, we shall, in keeping with the mathematical convention, not call them equivalent.

The first condition now means that if the states $\Phi^{(1)}$, $\Phi^{(2)} = S\Phi^{(1)}$ correspond to each other in one coördinate system, the states $D^{(1)}(L)\Phi^{(1)}$ and $D^{(2)}(L)\Phi^{(2)}$ correspond to each other also. We have then

$$D^{(2)}(L)\Phi^{(2)} = SD^{(1)}(L)\Phi^{(1)} = SD^{(1)}(L)S^{-1}\Phi^{(2)}. \tag{4a}$$

---

[7] E. Majorana, Nuovo Cim. *9*, 335, 1932.

[8] P. A. M. Dirac, Proc. Roy. Soc. A. *155*, 447, 1936; Al. Proca, J. de Phys. Rad. *7*, 347, 1936.

[9] Klein, Arkiv f. Matem. Astr. och Fysik, *25A*, No. 15, 1936. I am indebted to Mr. Darling for an interesting conversation on this paper.

As this shall hold for every $\Phi^{(2)}$, the existence of a unitary S which transforms $D^{(1)}$ into $D^{(2)}$ is the condition for the equivalence of these two representations. Equivalent representations are not considered to be really different and it will be sufficient to find one sample from every infinite class of equivalent representations.

If there is a closed linear manifold of states which is invariant under all Lorentz transformations, i.e. which contains $D(L)\psi$ if it contains ψ, the linear manifold perpendicular to this one will be invariant also. In fact, if φ belongs to the second manifold, $D(L)\varphi$ will be, on account of the unitary character of $D(L)$, perpendicular to $D(L)\psi'$ if ψ' belongs to the first manifold. However, $D(L^{-1})\psi$ belongs to the first manifold if ψ does and thus $D(L)\varphi$ will be orthogonal to $D(L)D(L^{-1})\psi = \omega\psi$ i.e. to all members of the first manifold and belong itself to the second manifold also. The original representation then "decomposes" into two representations, corresponding to the two linear manifolds. It is clear that, conversely, one can form a representation, by simply "adding" several other representations together, i.e. by considering as states linear combinations of the states of several representations and assume that the states which originate from different representations are perpendicular to each other.

Representations which are equivalent to sums of already known representations are not really new and, in order to master all representations, it will be sufficient to determine those, out of which all others can be obtained by "adding" a finite or infinite number of them together.

Two simple theorems shall be mentioned here which will be proved later (Sections 7A and 8C respectively). The first one refers to unitary representations of any closed group, the second to irreducible unitary representations of any (closed or open) group.

The representations of a closed group by unitary *operators* can be transformed into the sum of unitary representations with matrices of finite dimensions.

Given two non equivalent irreducible unitary representations of an arbitrary group. If the scalar product between the wave functions is invariant under the operations of the group, the wave functions belonging[28] to the first representation are orthogonal to all wave functions belonging to the second representation.

C. Classification of unitary representations according to von Neumann and Murray[10]

Given the operators $D(L)$ of a unitary representations, or a representation up to a factor, one can consider the algebra of these operators, i.e. all linear combinations

$$a_1 D(L_1) + a_2 D(L_2) + a_3 D(L_3) + \cdots$$

of the $D(L)$ and all limits of such linear combinations which are bounded operators. According to the properties of this representation algebra, three classes of unitary representations can be distinguished.

---

[10] F. J. Murray and J. v. Neumann, Ann. of Math. *37*, 116, 1936; J. v. Neumann, to be published soon.

The first class of *irreducible* representations has a representation algebra which contains all bounded operators, i.e. if ψ and φ are two arbitrary states, there is an operator A of the representation algebra for which $A\psi = \varphi$ and $A\psi' = 0$ if ψ' is orthogonal to ψ. It is clear that the center of the algebra contains only the unit operator and multiply thereof. In fact, if C is in the center one can decompose $C\psi = \alpha\psi + \psi'$ so that ψ' shall be orthogonal to ψ. However, ψ' must vanish since otherwise C would not commute with the operator which leaves ψ invariant and transforms every function orthogonal to it into 0. For similar reasons, α must be the same for all ψ. For irreducible representations there is no closed linear manifold of states, (excepting the manifold of all states) which is invariant under all Lorentz transformations. In fact, according to the above definition, a φ' arbitrarily close to any φ can be represented by a finite linear combination

$$a_1 D(L_1)\psi + a_2 D(L_2)\psi + \cdots + a_n D(L_n)\psi.$$

Hence, a closed linear invariant manifold contains every state if it contains one. This is, in fact, the more customary definition for irreducible representations and the one which will be used subsequently. It is well known that all finite dimensional representations are sums of irreducible representations. This is not true,[10] in general, in an infinite number of dimensions.

The second class of representations will be called *factorial*. For these, the center of the representation algebra still contains only multiples of the unit operator. Clearly, the irreducible representations are all factorial, but not conversely. For finite dimensions, the factorial representations may contain one irreducible representation several times. This is also possible in an infinite number of dimensions, but in addition to this, there are the "continuous" representations of Murray and von Neumann.[10] These are not irreducible as there are invariant linear manifolds of states. On the other hand, it is impossible to carry the decomposition so far as to obtain as parts only irreducible representations. In all the examples known so far, the representations into which these continuous representations can be decomposed, are equivalent to the original representation.

The third class contains all possible unitary representations. In a finite number of dimensions, these can be decomposed first into factorial representations, and these, in turn, in irreducible ones. Von Neumann[10] has shown that the first step still is possible in infinite dimensions. We can assume, therefore, from the outset that we are dealing with factorial representations.

In the theory of representations of finite dimensions, it is sufficient to determine only the irreducible ones, all others are equivalent to sums of these. Here, it will be necessary to determine all factorial representations. Having done that, we shall know from the above theorem of von Neumann, that all representations are equivalent to finite or infinite sums of factorial representations.

It will be one of the results of the detailed investigation that the inhomogeneous Lorentz group has no "continuous" representations, all representations

can be decomposed into irreducible ones. Thus the work of Majorana and Dirac appears to be justified from this point of view a posteriori.

D. Classification of unitary representations from the point of view of infinitesimal operators

The existence of an infinitesimal operator of a continuous one parametric (cyclic, abelian) unitary group has been shown by Stone.[11] He proved that the operators of such a group can be written as $\exp(iHt)$ where H is a (bounded or unbounded) hermitean operator and t is the group parameter. However, the Lorentz group has many one parametric subgroups, and the corresponding infinitesimal operators $H_1, H_2, \cdots$ are all unbounded. For every H_i an everywhere dense set of functions φ can be found such that $H_i\varphi$ can be defined. It is not clear, however, that an everywhere dense set can be found, to all members of which every H can be applied. In fact, it is not clear that one such φ can be found.

Indeed, it may be interesting to remark that for an irreducible representation the existence of one function φ to which all infinitesimal operators can be applied, entails the existence of an everywhere dense set of such functions. This again has the consequence that one can operate with infinitesimal operators to a large extent in the usual way.

PROOF: Let $Q(t)$ be a one parametric subgroup such that $Q(t)Q(t') = Q(t+t')$. If the infinitesimal operator of all subgroups can be applied to φ, the

$$\lim_{t=0} t^{-1}(Q(t) - 1)\varphi \tag{5}$$

exists. It follows, then, that the infinitesimal operators can be applied to $R\varphi$ also where R is an arbitrary operator of the representation: Since $R^{-1}Q(t)\,R$ is also a one parametric subgroup

$$\lim_{t=0} t^{-1}(R^{-1}Q(t)R - 1)\varphi = \lim_{t=0} R^{-1}\cdot t^{-1}(Q(t) - 1)R\varphi$$

also exists and hence also (R is unitary)

$$\lim_{t=0} t^{-1}(Q(t) - 1)R\varphi.$$

Every infinitesimal operator can be applied to $R\varphi$ if they all can be applied to φ, and the same holds for sums of the kind

$$a_1R_1\varphi + a_2R_2\varphi + \cdots + a_nR_n\varphi. \tag{6}$$

These form, however, an everywhere dense set of functions if the representation is irreducible.

If the representation is not irreducible, one can consider the set N_0 of such wave functions to which every infinitesimal operator can be applied. This set is

[11] M. H. Stone, Proc. Nat. Acad. *16*, 173, 1930, Ann. of Math. *33*, 643, 1932, also J. v. Neumann, ibid, *33*, 567, 1932.

clearly linear and, according to the previous paragraph, invariant under the operations of the group (i.e. contains every $R\varphi$ if it contains φ). The same holds for the closed set N generated by N_0 and also of the set P of functions which are perpendicular to all functions of N. In fact, if φ_p is perpendicular to all φ_n of N, it is perpendicular also to all $R^{-1}\varphi_n$ and, for the unitary character of R, the $R\varphi_p$ is perpendicular to all φ_n, i.e. is also contained in the set P.

We can decompose thus, by a unitary transformation, every unitary representation into a "normal" and a "pathological" part. For the former, there is an everywhere dense set of functions, to which all infinitesimal operators can be applied. There is no single wave functions to which all infinitesimal operators of a "pathological" representation could be applied.

According to Murray and von Neumann, if the original representation was factorial, all representations into which it can be decomposed will be factorial also. Thus every representation is equivalent to a sum of factorial representations, part of which is "normal," the other part "pathological."

It will turn out again that the inhomogeneous Lorentz group has no pathological representations. Thus this assumption of Majorana and Dirac also will be justified a posteriori. Every unitary representation of the inhomogenous Lorentz group can be decomposed into normal irreducible representations. It should be stated, however, that the representations in which the unit operator corresponds to every translation have not been determined to date (cf. also section 3, end). Hence, the above statements are not proved for these representations, which are, however, more truly representations of the homogeneous Lorentz group, than of the inhomogeneous group.

While all these points may be of interest to the mathematician only, the new representation of the Lorentz group which will be described in section 7 may interest the physicist also. It describes a particle with a continuous spin.

Acknowledgement. The subject of this paper was suggested to me as early as 1928 by P. A. M. Dirac who realised even at that date the connection of representations with quantum mechanical equations. I am greatly indebted to him also for many fruitful conversations about this subject, especially during the years 1934/35, the outgrowth of which the present paper is.

I am indebted also to J. v. Neumann for his help and friendly advice.

3. Summary of Ensuing Sections

Section 4 will be devoted to the definition of the inhomogeneous Lorentz group and the theory of characteristic values and characteristic vectors of a homogeneous (ordinary) Lorentz transformation. The discussion will follow very closely the corresponding, well-known theory of the group of motions in ordinary space and the theory of characteristic values of orthogonal transformations.[12] It will contain only a straightforward generalization of the methods usually applied in those discussions.

[12] Cf. e.g. E. Wigner, l.c. Chapter III. O. Veblen and J. W. Young, *Projective Geometry*, Boston 1917. Vol. 2, especially Chapter VII.

In Section 5, it will be proved that one can determine the physically meaningless constants in the $D(L)$ in such a way that instead of (3a) the more special equation

$$D(L_1)D(L_2) = \pm D(L_1L_2) \tag{7}$$

will be valid. This means that instead of a representation up to a factor, we can consider representations up to the sign. For the case that either L_1 or L_2 is a pure translation, Dirac[13] has given a proof of (7) using infinitesimal operators. A consideration very similar to his can be carried out, however, also using only finite transformations.

For representations with a finite number of dimensions (corresponding to an only finite number of linearly independent states), (7) could be proved also if both L_1 and L_2 are homogeneous Lorentz transformations, by a straightforward application of the method of Weyl and Schreier.[14] However, the Lorentz group has no finite dimensional representation (apart from the trivial one in which the unit operation corresponds to every L). Thus the method of Weyl and Schreier cannot be applied. Its first step is to normalize the indeterminate constants in every matrix $D(L)$ in such a way that the determinant of $D(L)$ becomes 1. No determinant can be defined for general unitary operators.

The method to be employed here will be to decompose every L into a product of two involutions $L = MN$ with $M^2 = N^2 = 1$. Then $D(M)$ and $D(N)$ will be normalized so that their squares become unity and $D(L) = D(M)D(N)$ set. It will be possible, then, to prove (7) without going back to the topology of the group.

Sections 6, 7, and 8 will contain the determination of the representations. The pure translations form an invariant subgroup of the whole inhomogeneous Lorentz group and Frobenius' method[15] will be applied in Section 6 to build up the representations of the whole group out of representations of the subgroup, by means of a "little group." In Section 6, it will be shown on the basis of an as yet unpublished work[24] of J. v. Neumann that there is a characteristic (invariant) set of "momentum vectors" for every irreducible representation. The irreducible representations of the Lorentz group will be divided into four classes. The momentum vectors of the

1*st class* are time-like,
2*nd class* are null-vectors, but not all their components will be zero,
3*rd class* vanish (i.e., all their components will be zero),
4*th class* are space-like.

Only the first two cases will be considered in Section 7, although the last case

---

[13] P. A. M. Dirac, mimeographed notes of lectures delivered at Princeton University, 1934/35, page 5a.

[14] H. Weyl, Mathem. Zeits. *23*, 271; *24*, 328, 377, 789, 1925; O. Schreier, Abhandl. Mathem. Seminar Hamburg, *4*, 15, 1926; *5*, 233, 1927.

[15] G. Frobenius, Sitz. d. Kön. Preuss. Akad. p. 501, 1898, I. Schur, ibid, p. 164, 1906; F. Seitz, Ann. of Math. *37*, 17, 1936.

may be the most interesting from the mathematical point of view. I hope to return to it in another paper. I did not succeed so far in giving a complete discussion of the 3rd class. (All these restrictions appear in the previous treatments also.)

In Section 7, we shall find again all known representations of the inhomogeneous Lorentz group (i.e., all known Lorentz invariant equations) and two new sets.

Sections 5, 6, 7 will deal with the "restricted Lorentz group" only, i.e. Lorentz transformations with determinant 1 which do not reverse the direction of the time axis. In section 8, the representations of the extended Lorentz group will be considered, the transformations of which are not subject to these conditions.

4. Description of the Inhomogeneous Lorentz Group

A.

An inhomogeneous Lorentz transformation $L = (a, \Lambda)$ is the product of a translation by a real vector a

$$x_i' = x_i + a_i \qquad (i = 1, 2, 3, 4) \tag{8}$$

and a homogeneous Lorentz transformation Λ with real coefficients

$$x_i' = \sum_{k=1}^{4} \Lambda_{ik} x_k . \tag{9}$$

The translation shall be performed after the homogeneous transformation. The coefficients of the homogeneous transformation satisfy three conditions: (1) They are real and Λ leaves the indefinite quadratic form $-x_1^2 - x_2^2 - x_3^2 + x_4^2$ invariant:

$$\Lambda F \Lambda' = F \tag{10}$$

where the prime denotes the interchange of rows and columns and F is the diagonal matrix with the diagonal elements $-1, -1, -1, +1$.—(2) The determinant $|\Lambda_{ik}| = 1$ and—(3) $\Lambda_{44} > 0$.

We shall denote the Lorentz-hermitean product of two vectors x and y by

$$\{x, y\} = -x_1^* y_1 - x_2^* y_2 - x_3^* y_3 + x_4^* y_4 . \tag{11}$$

(The star denotes the conjugate imaginary.) If $\{x, x\} < 0$ the vector x is called space-like, if $\{x, x\} = 0$, it is a null vector, if $\{x, x\} > 0$, it is called time-like. A real time-like vector lies in the positive light cone if $x_4 > 0$; it lies in the negative light cone if $x_4 < 0$. Two vectors x and y are called orthogonal if $\{x, y\} = 0$.

On account of its linear character a homogeneous Lorentz transformation is completely defined if Λv is given for four linearly independent vectors $v^{(1)}$, $v^{(2)}$, $v^{(3)}$, $v^{(4)}$.

From (11) and (10) it follows that $\{v, w\} = \{\Lambda v, \Lambda w\}$ for every pair of vectors v, w. This will be satisfied for every pair if it is satisfied for all pairs $v^{(i)}$, $v^{(k)}$

of four linearly independent vectors. The reality condition is satisfied if $(\Lambda v^{(i)})^* = \Lambda(v^{(i)*})$ holds for four such vectors.

The scalar product of two vectors x and y is positive if both lie in the positive light cone or both in the negative light cone. It is negative if one lies in the positive, the other in the negative light cone. Since both x and y are time-like $|x_4|^2 > |x_1|^2 + |x_2|^2 + |x_3|^2$; $|y_4|^2 > |y_1|^2 + |y_2|^2 + |y_3|^2$. Hence, by Schwarz's inequality $|x_4^* y_4| > |x_1^* y_1 + x_2^* y_2 + x_3^* y_3|$ and the sign of the scalar product of two real time-like vectors is determined by the product of their time components.

A time-like vector is transformed by a Lorentz transformation into a time-like vector. Furthermore, on account of the condition $\Lambda_{44} > 0$, the vector $v^{(0)}$ with the components 0, 0, 0, 1 remains in the positive light cone, since the fourth component of $\Lambda v^{(0)}$ is Λ_{44}. If $v^{(1)}$ is another vector[16] in the positive light cone $\{v^{(1)}, v^{(0)}\} > 0$ and hence also $\{\Lambda v^{(1)}, \Lambda v^{(0)}\} > 0$ and $\Lambda v^{(1)}$ is in the positive light cone also. The third condition for a Lorentz transformation can be formulated also as the requirement that every vector in (or on) the positive light cone shall remain in (or, respectively, on) the positive light cone.

This formulation of the third condition shows that the third condition holds for the product of two homogeneous Lorentz transformations if it holds for both factors. The same is evident for the first two conditions.

From $\Lambda F \Lambda' = F$ one obtains by multiplying with Λ^{-1} from the left and $\Lambda'^{-1} = (\Lambda^{-1})'$ from the right $F = \Lambda^{-1} F (\Lambda^{-1})'$ so that the reciprocal of a homogeneous Lorentz transformation is again such a transformation. The homogeneous Lorentz transformations form a group, therefore.

• One easily calculates that the product of two inhomogeneous Lorentz transformations (b, M) and (c, N) is again an inhomogeneous Lorentz transformation (a, Λ)

$$(b, M)(c, N) = (a, \Lambda) \tag{12}$$

where

$$\Lambda_{ik} = \sum_j M_{ij} N_{jk}; \qquad a_i = b_i + \sum_j M_{ij} c_j, \tag{12a}$$

or, somewhat shorter

$$\Lambda = MN; \qquad a = b + Mc. \tag{12b}$$

B. Theory of characteristic values and characteristic vectors of a homogeneous Lorentz transformation

Linear homogeneous transformations are most simply described by their characteristic values and vectors. Before doing this for the homogeneous Lorentz group, however, we shall need two rules about orthogonal vectors.

[16] Wherever a confusion between vectors and vector components appears to be possible, upper indices will be used for distinguishing different vectors and lower indices for denoting the components of a vector.

[1] *If* $\{v, w\} = 0$ *and* $\{v, v\} > 0$, *then* $\{w, w\} < 0$; *if* $\{v, w\} = 0$, $\{v, v\} = 0$, *then* w *is either space-like, or parallel to* v (*either* $\{w, w\} < 0$, *or* $w = cv$).

PROOF:

$$v_4^* w_4 = v_1^* w_1 + v_2^* w_2 + v_3^* w_3 . \tag{13}$$

By Schwarz's inequality, then

$$|v_4|^2 |w_4|^2 \leqq (|v_1|^2 + |v_2|^2 + |v_3|^2)(|w_1|^2 + |w_2|^2 + |w_3|^2). \tag{14}$$

For $|v_4|^2 > |v_1|^2 + |v_2|^2 + |v_3|^2$ it follows that $|w_4|^2 < |w_1|^2 + |w_2|^2 + |w_3|^2$. If $|v_4|^2 = |v_1|^2 + |v_2|^2 + |v_3|^2$ the second inequality still follows if the inequality sign holds in (14). The equality sign can hold only, however, if the first three components of the vectors v and w are proportional. Then, on account of (13) and both being null vectors, the fourth components are in the same ratio also.

[2] *If four vectors* $v^{(1)}$, $v^{(2)}$, $v^{(3)}$, $v^{(4)}$ *are mutually orthogonal and linearly independent, one of them is time-like, three are space-like.*

PROOF: It follows from the previous paragraph that only one of four mutually orthogonal, linearly independent vectors can be time-like or a null vector. It remains to be shown therefore only that one of them *is* time-like. Since they are linearly independent, it is possible to express by them any time-like vector

$$v^{(l)} = \sum_{k=1}^{4} \alpha_k v^{(k)}.$$

The scalar product of the left side of this equation with itself is positive and therefore

$$\left\{ \sum_k \alpha_k v^{(k)}, \quad \sum_k \alpha_k v^{(k)} \right\} > 0$$

or

$$\sum_k |\alpha_k|^2 \{v^{(k)}, v^{(k)}\} > 0 \tag{15}$$

and one $\{v^{(k)}, v^{(k)}\}$ must be positive. Four mutually orthogonal vectors are not necessarily linearly independent, because a null vector is perpendicular to itself. The linear independence follows, however, if none of the four is a null vector.

We go over now to the characteristic values λ of Λ. These make the determinant $|\Lambda - \lambda 1|$ of the matrix $\Lambda - \lambda 1$ vanish.

[3] *If* λ *is a characteristic value,* λ^*, λ^{-1} *and* λ^{*-1} *are characteristic values also.*

PROOF: For λ^* this follows from the fact that Λ is real. Furthermore, from $|\Lambda - \lambda 1| = 0$ also $|\Lambda' - \lambda 1| = 0$ follows, and this multiplied by the determinants of ΛF and F^{-1} gives

$$|\Lambda F| \cdot |\Lambda' - \lambda 1| \cdot |F|^{-1} = |\Lambda F \Lambda' F^{-1} - \lambda \Lambda| = |1 - \lambda \Lambda| = 0,$$

so that λ^{-1} is a characteristic value also.

[4] *The characteristic vectors* v_1 *and* v_2 *belonging to two characteristic values* λ_1 *and* λ_2 *are orthogonal if* $\lambda_1^* \lambda_2 \neq 1$.

PROOF:

$$\{v_1, v_2\} = \{\Lambda v_1, \Lambda v_2\} = \{\lambda_1 v_1, \lambda_2 v_2\} = \lambda_1^* \lambda_2 \{v_1, v_2\}.$$

Thus if $\{v_1, v_2\} \neq 0$, $\lambda_1^* \lambda_2 = 1$.

[5] *If the modulus of a characteristic value* λ *is* $|\lambda| \neq 1$, *the corresponding characteristic vector* v *is a null vector and* λ *itself real and positive.*

From $\{v, v\} = \{\Lambda v, \Lambda v\} = |\lambda|^2 \{v, v\}$ the $\{v, v\} = 0$ follows immediately for $|\lambda| \neq 1$. If λ were complex, λ^* would be a characteristic value also. The characteristic vectors of λ and λ^* would be two different null vectors and, because of [4], orthogonal to each other. This is impossible on account of [1]. Thus λ is real and v a real null vector. Then, on account of the third condition for a homogeneous Lorentz transformation, λ must be positive.

[6] *The characteristic value* λ *of a characteristic vector* v *of length null is real and positive.*

If λ were not real, λ^* would be a characteristic value also. The corresponding characteristic vector v^* would be different from v, a null vector also, and perpendicular to v on account of [4]. This is impossible because of [1].

[7] *The characteristic vector* v *of a complex characteristic value* λ (*the modulus of which is* 1 *on account of* [5]) *is space-like*: $\{v, v\} < 0$.

PROOF: λ^* is a characteristic value also, the corresponding characteristic vector is v^*. Since $(\lambda^*)^* \lambda = \lambda^2 \neq 1$, $\{v^*, v\} = 0$. Since they are different, at least one is space-like. On account of $\{v, v\} = \{v^*, v^*\}$ both are space-like. If all four characteristic values were complex and the corresponding characteristic vectors linearly independent (which is true except if Λ has elementary divisors) we should have four space-like, mutually orthogonal vectors. This is impossible, on account of [2]. Hence

[8] *There is not more than one pair of conjugate complex characteristic values, if* Λ *has no elementary divisors. Similarly, under the same condition, there is not more than one pair* λ, λ^{-1} *of characteristic values whose modulus is different from* 1. Otherwise their characteristic vectors would be orthogonal, which they cannot be, being null vectors.

For homogeneous Lorentz transformations which do not have elementary divisors, the following possibilities remain:

(a) There is a pair of complex characteristic values, their modulus is 1, on account of [5]

$$\lambda_1 = \lambda_2^* = \lambda_2^{-1}; \quad |\lambda_1| = |\lambda_2| = 1, \tag{16}$$

and also a pair of characteristic values λ_3, λ_4, the modulus of which is not 1. These must be real and positive:

$$\lambda_4 = \lambda_3^{-1}; \quad \lambda_3 = \lambda_3^* > 0. \tag{16a}$$

The characteristic vectors of the conjugate complex characteristic values are conjugate complex, perpendicular to each other and space-like so that they can be normalized to -1

$$v_1 = v_2^*; \quad \{v_1, v_2\} = \{v_1, v_1^*\} = 0 \tag{17}$$
$$\{v_1, v_1\} = \{v_2, v_2\} = -1$$

those of the real characteristic values are real null vectors, their scalar product can be normalized to 1

$$v_3 = v_3^* \qquad v_4 = v_4^* \qquad \{v_3, v_4\} = 1$$
$$\{v_3, v_3\} = \{v_4, v_4\} = 0. \tag{17a}$$

Finally, the former pair of characteristic vectors is perpendicular to the latter kind

$$\{v_1, v_3\} = \{v_1, v_4\} = \{v_2, v_3\} = \{v_2, v_4\} = 0. \tag{17b}$$

It will turn out that all the other cases in which Λ has no elementary divisor are special cases of (a).

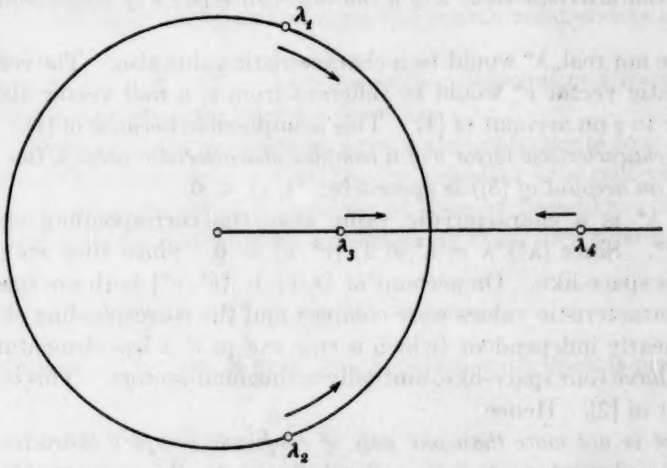


FIG. 1. Position of the characteristic values for the general case a) in the complex plane. In case b), λ_3 and λ_4 coincide and are equal 1; in case c), λ_1 and λ_2 coincide and are either $+1$ or -1. In case d) both pairs $\lambda_3 = \lambda_4 = 1$ and $\lambda_1 = \lambda_2 = \pm 1$ coincide.

(b) There is a pair of complex characteristic values λ_1, $\lambda_2 = \lambda_1^{-1} = \lambda_1^*$, $\lambda_1 \neq \lambda_1^*$, $|\lambda_1| = |\lambda_2| = 1$. No pair with $|\lambda_3| \neq 1$, however. Then on account of [8], still $\lambda_3 = \lambda_3^*$ which gives with $|\lambda_3| = 1$, $\lambda_3 = \pm 1$. Since the product $\lambda_1\lambda_2\lambda_3\lambda_4 = 1$, on account of the second condition for homogeneous Lorentz transformations, also $\lambda_4 = \lambda_3 = \pm 1$. The double characteristic value ± 1 has two linearly independent characteristic vectors v_3 and v_4 which can be assumed to be perpendicular to each other, $\{v_3, v_4\} = 0$. According to [2], one of the four characteristic vectors must be time-like and since those of λ_1 and λ_2 are space-like, the time-like one must belong to ± 1. This must be positive, therefore $\lambda_3 = \lambda_4 = 1$. Out of the time-like and space-like vectors $\{v_3, v_3\} = -1$ and $\{v_4, v_4\} = 1$, one can build two null vectors $v_4 + v_3$ and $v_4 - v_3$. Doing this, case (b) becomes the special case of (a) in which the real positive characteristic values become equal $\lambda_3 = \lambda_4^{-1} = 1$.

(c) All characteristic values are real; there is however one pair $\lambda_3 = \lambda_3^*$,

$\lambda_4 = \lambda_3^{-1}$, the modulus of which is not unity. Then $\{v_3, v_3\} = \{v_4, v_4\} = 0$ and $\lambda_3 > 0$ and one can conclude for λ_1 and λ_2, as before for λ_3 and λ_4 that $\lambda_1 = \lambda_2 = \pm 1$. This again is a special case of (a); here the two characteristic values of modulus 1 become equal.

(d) All characteristic values are real and of modulus 1. If all of them are $+1$, we have the unit matrix which clearly can be considered as a special case of (a). The other case is $\lambda_1 = \lambda_2 = -1$, $\lambda_3 = \lambda_4 = +1$. The characteristic vectors of λ_1 and λ_2 must be space-like, on account of the third condition for a homogeneous Lorentz transformation; they can be assumed to be orthogonal and normalized to -1. This is then a special case of (b) and hence of (a) also. The cases (a), (b), (c), (d) are illustrated in Fig. 1.

The cases remain to be considered in which Λ has an elementary divisor. We set therefore

$$\Lambda_e v_e = \lambda_e v_e\,; \qquad \Lambda_e w_e = \lambda_e w_e + v_e\,. \tag{18}$$

It follows from [5] that either $|\lambda_e| = 1$, or $\{v_e, v_e\} = 0$. We have $\{v_e, w_e\} = \{\Lambda_e v_e, \Lambda_e w_e\} = |\lambda_e|^2\{v_e, w_e\} + \{v_e, v_e\}$. From this equation

$$\{v_e, v_e\} = 0 \tag{19}$$

follows for $|\lambda_e| = 1$, so that (19) holds in any case. It follows then from [6] that λ_e is real, positive and v_e, w_e can be assumed to be real also. The last equation now becomes $\{v_e, w_e\} = \lambda_e^2\{v_e, w_e\}$ so that either $\lambda_e = 1$ or $\{v_e, w_e\} = 0$. Finally, we have

$$\{w_e, w_e\} = \{\Lambda_e w_e, \Lambda_e w_e\} = \lambda_e^2\{w_e, w_e\} + 2\lambda_e\{w_e, v_e\} + \{v_e, v_e\}.$$

This equation now shows that

$$\{w_e, v_e\} = 0 \tag{19a}$$

even if $\lambda_e = 1$. From (19), (19a) it follows that w_e is space-like and can be normalized to

$$\{w_e, w_e\} = -1. \tag{19b}$$

Inserting (19a) into the preceding equation we finally obtain

$$\lambda_e = 1. \tag{19c}$$

[9] *If Λ_e has an elementary divisor, all its characteristic roots are* 1.

From (19c) we see that the root of the elementary divisor is 1 and this is at least a double root. If Λ had a pair of characteristic values $\lambda_1 \neq 1$, $\lambda_2 = \lambda_1^{-1}$, the corresponding characteristic vectors v_1 and v_2 would be orthogonal to v_e and therefore space-like. On account of [5], then $|\lambda_1| = |\lambda_2| = 1$ and $\{v_1, v_2\} = 0$. Furthermore, from $\{w_e, v_1\} = \{\Lambda_e w_e, \Lambda_e v_1\} = \lambda_1\{w_e, v_1\} + \lambda_1\{v_e, v_1\}$ and from $\{v_e, v_1\} = 0$ also $\{w_e, v_1\} = 0$ follows. Thus all the four vectors v_1, v_2, v_e, w_e would be mutually orthogonal. This is excluded by [2] and (19).

Two cases are conceivable now. Either the fourfold characteristic root has only one characteristic vector, or there is in addition to v_e (at least) another characteristic vector v_1. In the former case four linearly independent vectors v_e, w_e, z_e, x_e could be found such that

$$\Lambda_e v_e = v_e \qquad \Lambda_e w_e = w_e + v_e$$
$$\Lambda_e z_e = z_e + w_e \qquad \Lambda_e x_e = x_e + z_e.$$

However $\{v_e, x_e\} = \{\Lambda_e v_e, \Lambda_e x_e\} = \{v_e, x_e\} + \{v_e, z_e\}$ from which $\{v_e, z_e\} = 0$ follows. On the other hand

$$\{w_e, z_e\} = \{\Lambda_e w_e, \Lambda_e z_e\} = \{w_e, z_e\} + \{w_e, w_e\} + \{v_e, z_e\} + \{v_e, w_e\}.$$

This gives with (19a) and (19b) $\{v_e, z_e\} = 1$ so that this case must be excluded.

(e) There is thus a vector v_1 so that in addition to (18)

$$\Lambda_e v_1 = v_1 \tag{18a}$$

holds. From $\{w_e, v_1\} = \{\Lambda_e w_e, \Lambda_e v_1\} = \{w_e, v_1\} + \{v_e, v_1\}$ follows

$$\{v_e, v_1\} = 0. \tag{19d}$$

The equations (18), (18a) will remain unchanged if we add to w_e and v_1 a multiple of v_e. We can achieve in this way that the fourth components of both w_e and v_1 vanish. Furthermore, v_1 can be normalized to -1 and added to w_e also with an arbitrary coefficient, to make it orthogonal to v_1. Hence, we can assume that

$$v_{14} = w_{e4} = 0; \quad \{v_1, v_1\} = -1; \quad \{w_e, v_1\} = 0. \tag{19e}$$

We can finally define the null vector z_e to be orthogonal to w_e and v_1 and have a scalar product 1 with v_e

$$\{z_e, z_e\} = \{z_e, w_e\} = \{z_e, v_1\} = 0; \quad \{z_e, v_e\} = 1. \tag{19f}$$

Then the null vectors v_e and z_e represent the momenta of two light beams in opposite directions. If we set $\Lambda_e z_e = a v_e + b w_e + c z_e + d v_1$ the conditions $\{z_e, v\} = \{\Lambda_e z_e, \Lambda_e v\}$ give, if we set for v the vectors v_e, w_e, z_e, v_1 the conditions $c = 1; b = c; 2ac - b^2 - d^2 = 0; d = 0$. Hence

$$\begin{aligned} \Lambda_e v_e &= v_e \qquad \Lambda_e w_e = w_e + v_e \\ \Lambda_e v_1 &= v_1 \qquad \Lambda_e z_e = z_e + w_e + \tfrac{1}{2} v_e. \end{aligned} \tag{20}$$

A Lorentz transformation with an elementary divisor can be best characterized by the null vector v_e which is invariant under it and the space part of which forms with the two other vectors w_e and v_1 three mutually orthogonal vectors in ordinary space. The two vectors w_e and v_1 are normalized, v_1 is invariant under Λ_e while the vector v_e is added to w_e upon application of Λ_e. The result of the application of Λ_e to a vector which is linearly independent of v_e, w_e and v_1 is, as we saw, already determined by the expressions for $\Lambda_e v_e$, $\Lambda_e w_e$ and $\Lambda_e v_1$.

The $\Lambda_e(\gamma)$ which have the invariant null vector v_e and also w_e (and hence also

v_1) in common and differ only by adding to w_ε different multiples γv_ε of v_ε, form a cyclic group with $\gamma = 0$, the unit transformation as unity:

$$\Lambda_\varepsilon(\gamma)\Lambda_\varepsilon(\gamma') = \Lambda_\varepsilon(\gamma + \gamma').$$

The Lorentz transformation $M(\alpha)$ which leaves v_1 and w_ε invariant but replaces v_ε by αv_ε (and z_ε by $\alpha^{-1} z_\varepsilon$) has the property of transforming $\Lambda_\varepsilon(\gamma)$ into

$$M(\alpha)\Lambda_\varepsilon(\gamma)M(\alpha)^{-1} = \Lambda_\varepsilon(\alpha\gamma). \qquad (+)$$

An example of $\Lambda_\varepsilon(\gamma)$ and $M(\alpha)$ is

$$\Lambda_\varepsilon(\gamma) = \begin{Vmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & \gamma & \gamma \\ 0 & -\gamma & 1 - \tfrac{1}{2}\gamma^2 & -\tfrac{1}{2}\gamma^2 \\ 0 & \gamma & \tfrac{1}{2}\gamma^2 & 1 + \tfrac{1}{2}\gamma^2 \end{Vmatrix};$$

$$M(\alpha) = \begin{Vmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \tfrac{1}{2}(\alpha + \alpha^{-1}) & \tfrac{1}{2}(\alpha - \alpha^{-1}) \\ 0 & 0 & \tfrac{1}{2}(\alpha - \alpha^{-1}) & \tfrac{1}{2}(\alpha + \alpha^{-1}) \end{Vmatrix}.$$

These Lorentz transformations play an important rôle in the representations with space like momentum vectors.

A behavior like (+) is impossible for finite unitary matrices because the characteristic values of $M(\alpha)^{-1}\Lambda_\varepsilon(\gamma)M(\alpha)$ and $\Lambda_\varepsilon(\gamma)$ are the same—those of $\Lambda_\varepsilon(\gamma\alpha) = \Lambda_\varepsilon(\gamma)^\alpha$ the α^{th} powers of those of $\Lambda_\varepsilon(\gamma)$. This shows very simply that the Lorentz group has no true unitary representation in a finite number of dimensions.

C. Decomposition of a homogeneous Lorentz transformation into rotations and an acceleration in a given direction

The homogeneous Lorentz transformation is, from the point of view of the physicist, a transformation to a uniformly moving coördinate system, the origin of which coincided at $t = 0$ with the origin of the first coördinate system. One can, therefore, first perform a rotation which brings the direction of motion of the second system into a given direction—say the direction of the third axis—and impart it a velocity in this direction, which will bring it to rest. After this, the two coördinate systems can differ only in a rotation. This means that every homogeneous Lorentz transformation can be decomposed in the following way[17]

$$\Lambda = RZS \qquad (21)$$

[17] Cf. e.g. L. Silberstein, *The Theory of Relativity*, London 1924, p. 142.

where R and S are pure rotations, (i.e. $R_{ik} = R_{ki} = S_{ik} = S_{ki} = 0$ for $i \neq 4$ and $R_{44} = S_{44} = 1$, also $R' = R^{-1}$, $S' = S^{-1}$) and Z is an acceleration in the direction of the third axis, i.e.

$$Z = \begin{Vmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & a & b \\ 0 & 0 & b & a \end{Vmatrix}$$

with $a^2 - b^2 = 1$, $a > b > 0$. The decomposition (21) is clearly not unique. It will be shown, however, that Z is uniquely determined, i.e. the same in every decomposition of the form (21).

In order to prove this mathematically, we chose R so that in $R^{-1}\Lambda = I$ the first two components in the fourth column $I_{14} = I_{24} = 0$ become zero: R^{-1} shall bring the vector with the components Λ_{14}, Λ_{24}, Λ_{34} into the third axis. Then we take $I_{34} = (\Lambda_{14}^2 + \Lambda_{24}^2 + \Lambda_{34}^2)^{\frac{1}{2}}$ and $I_{44} = \Lambda_{44}$ for b and a to form Z; they satisfy the equation $I_{44}^2 - I_{34}^2 = 1$. Hence, the first three components of the fourth column of $J = Z^{-1}I = Z^{-1}R^{-1}\Lambda$ will become zero and $J_{44} = 1$, because of $J_{44}^2 - J_{14}^2 - J_{24}^2 - J_{34}^2 = 1$. Furthermore, the first three components of the fourth row of J will vanish also, on account of $J_{44}^2 - J_{41}^2 - J_{42}^2 - J_{43}^2 = 1$, i.e. $J = S = Z^{-1}R^{-1}\Lambda$ is a pure rotation. This proves the possibility of the decomposition (21).

The trace of $\Lambda\Lambda' = RZ^2R^{-1}$ is equal to the trace of Z^2, i.e. equal to $2a^2 + 2b^2 + 2 = 4a^2 = 4b^2 + 4$ which shows that the a and b of Z are uniquely determined. In particular $a = 1$, $b = 0$ and Z the unit matrix if $\Lambda\Lambda' = 1$, i.e. Λ a pure rotation.

It is easy to show now that the group space of the homogeneous Lorentz transformations is only doubly connected. If a continuous series $\Lambda(t)$ of homogeneous Lorentz transformations is given, which is unity both for $t = 0$ and for $t = 1$, we can decompose it according to (21)

$$\Lambda(t) = R(t)Z(t)S(t). \tag{21a}$$

It is also clear from the foregoing, that $R(t)$ can be assumed to be continuous in t, except for values of t, for which $\Lambda_{14} = \Lambda_{24} = \Lambda_{34} = 0$, i.e. for which Λ is a pure rotation. Similarly, $Z(t)$ will be continuous in t and this will hold even where $\Lambda(t)$ is a pure rotation. Finally, $S = Z^{-1}R^{-1}\Lambda$ will be continuous also, except where $\Lambda(t)$ is a pure rotation.

Let us consider now the series of Lorentz transformations

$$\Lambda_s(t) = R(t)Z(t)^sS(t) \tag{21b}$$

where the b of $Z(t)^s$ is s times the b of $Z(t)$. By decreasing s from 1 to 0 we continuously deform the set $\Lambda_1(t) = \Lambda(t)$ of Lorentz transformations into a set of rotations $\Lambda_0(t) = R(t)S(t)$. Both the beginning $\Lambda_0(0) = 1$ and the end $\Lambda_s(1) = 1$ of the set remain the unit matrix and the sets $\Lambda_s(t)$ remain continuous in t for

all values of s. This last fact is evident for such t for which $\Lambda(t)$ is not a rotation: for such t all factors of (21b) are continuous. But it is true also for t_0 for which $\Lambda(t_0)$ is a rotation, and for which, hence $Z(t_0) = 1$ and $\Lambda_s(t_0) = \Lambda_1(t_0) = \Lambda(t_0)$. As $Z(t)$ is everywhere continuous, there will be a neighborhood of t_0 in which $Z(t)$ and hence also $Z(t)^s$ is arbitrarily close to the unit matrix. In this neighborhood $\Lambda_s(t) = \Lambda(t)$. $S(t)^{-1}Z(t)^{-1}Z(t)^s\ S(t)$ is arbitrarily close to $\Lambda(t)$; and, if the neighborhood is small enough, this is arbitrarily close to $\Lambda(t_0) = \Lambda_s(t_0)$.

Thus (21b) replaces the continuous set $\Lambda(t)$ of Lorentz transformations by a continuous set of rotations. Since these form an only doubly connected manifold, the manifold of Lorentz transformations can not be more than doubly connected. The existence of a two valued representation[18] shows that it is actually doubly and not simply connected.

We can form a new group[14] from the Lorentz group, the elements of which are the elements of the Lorentz group, together with a way $\Lambda(t)$, connecting $\Lambda(1) = \Lambda$ with the unity $\Lambda(0) = E$. However, two ways which can be continuously deformed into each other are not considered different. The product of the element "Λ with the way $\Lambda(t)$" with the element "I with the way $I(t)$" is the element ΛI with the way which goes from E along $\Lambda(t)$ to Λ and hence along $\Lambda I(t)$ to ΛI. Clearly, the Lorentz group is isomorphic with this group and two elements (corresponding to the two essentially different ways to Λ) of this group correspond to one element of the Lorentz group. It is well known,[18] that this group is holomorphic with the group of unimodular complex two dimensional transformations.

Every continuous representation of the Lorentz group "up to the sign" is a singlevalued, continuous representation of this group. The transformation which corresponds to "Λ with the way $\Lambda(t)$" is that $d(\Lambda)$ which is obtained by going over from $d(E) = d(\Lambda(0)) = 1$ continuously along $d(\Lambda(t))$ to $d(\Lambda(1)) = d(\Lambda)$.

D. The homogeneous Lorentz group is simple

It will be shown, first, that an invariant subgroup of the homogeneous Lorentz group contains a rotation (i.e. a transformation which leaves x_4 invariant).— We can write an arbitrary element of the invariant subgroup in the form RZS of (21). From its presence in the invariant subgroup follows that of $S \cdot RZS \cdot S^{-1} = SRZ = TZ$. If X_π is the rotation by π about the first axis, $X_\pi ZX_\pi = Z^{-1}$ and $X_\pi TZX_\pi^{-1} = X_\pi TX_\pi X_\pi ZX_\pi = X_\pi TX_\pi Z^{-1}$ is contained in the invariant subgroup also and thus the transform of this with Z, i.e. $Z^{-1}X_\pi TX_\pi$ also. The product of this with TZ is $TX_\pi TX_\pi$ which leaves x_4 invariant. If $TX_\pi TX_\pi = 1$ we can take $TY_\pi TY_\pi$. If this is the unity also, $TX_\pi TX_\pi = TY_\pi TY_\pi$ and T commutes with $X_\pi Y_\pi$, i.e. is a rotation about the third axis. In this case the

---

[18] Cf. H. Weyl, *Gruppentheorie und Quantenmechanik*, 1st. ed. Leipzig 1928, pages 110–114, 2nd ed. Leipzig 1931, pages 130–133. It may be interesting to remark that essentially the same isomorphism has been recognized already by L. Silberstein, l.c. pages 148–157.

space like (complex) characteristic vectors of TZ lie in the plane of the first two coördinate axes. Transforming TZ by an acceleration in the direction of the first coördinate axis we obtain a new element of the invariant subgroup for which the space like characteristic vector will have a not vanishing fourth component. Taking this for RZS we can transform it with S again to obtain a new $SRZ = TZ$. However, since S leaves x_4 invariant, the fourth component of the space like characteristic vectors of this TZ will not vanish and we can obtain from it by the procedure just described a rotation which must be contained in the invariant subgroup.

It remains to be shown that an invariant subgroup which contains a rotation, contains the whole homogeneous Lorentz group. Since the three-dimensional rotation group is simple, all rotations must be contained in the invariant subgroup. Thus the rotation by π around the first axis X_π and also its transform with Z and also

$$ZX_\pi Z^{-1} \cdot X_\pi = Z \cdot X_\pi Z^{-1} X_\pi = Z^2$$

is contained in the invariant subgroup. However, the general acceleration in the direction of the third axis can be written in this form. As all rotations are contained in the invariant subgroup also, (21) shows that this holds for all elements of the homogeneous Lorentz group.

It follows from this that the homogeneous Lorentz group has apart from the representation with unit matrices only true representations. It follows then from the remark at the end of part B, that these have all infinite dimensions. This holds even for the two-valued representations to which we shall be led in Section 5 equ. (52D), as the group elements to which the positive or negative unit matrix corresponds must form an invariant subgroup also, and because the argument at the end of part B holds for two-valued representations also. One easily sees furthermore from the equations (52B), (52C) that it holds for the inhomogeneous Lorentz group equally well.

5. Reduction of Representations Up to a Factor to Two-Valued Representations

The reduction will be effected by giving each unitary transformation, which is defined by the physical content of the theory and the consideration of reference only up to a factor of modulus unity, a "phase," which will leave only the sign of the representation operators undetermined. The unitary operator corresponding to the translation a will be denoted by $T(a)$, that to the homogeneous Lorentz transformation Λ by $d(\Lambda)$. To the general inhomogeneous Lorentz transformation then $D(a, \Lambda) = T(a)d(\Lambda)$ will correspond. Instead of the relations (12), we shall use the following ones.

$$T(a)T(b) = \omega(a, b)T(a + b) \tag{22B}$$

$$d(\Lambda)T(a) = \omega(\Lambda, a)T(\Lambda a)d(\Lambda) \tag{22C}$$

$$d(\Lambda)d(I) = \omega(\Lambda, I)d(\Lambda I). \tag{22D}$$

The ω are numbers of modulus 1. They enter because the multiplication rules (12) hold for the representatives only up to a factor. Otherwise, the relations (22) are consequences of (12) and can in their return replace (12). We shall replace the $T(a)$, $d(\Lambda)$ by $\Omega(a)T(a)$ and $\Omega(\Lambda)d(\Lambda)$ respectively, for which equations similar to (22) hold, however with

$$\omega(a, b) = 1; \quad \omega(\Lambda, a) = 1; \quad \omega(\Lambda, I) = \pm 1. \tag{22'}$$

A.

It is necessary, first, to show that the undetermined factors in the representation $D(L)$ can be assumed in such a way that the $\omega(a, b)$, $\omega(\Lambda, a)$, $\omega(\Lambda, I)$ become —apart from regions of lower dimensionality—continuous functions of their arguments. This is a consequence of the continuous character of the representation and shall be discussed first.

(a) From the point of view of the physicist, the natural definition of the continuity of a representation up to a factor is as follows. The neighborhood δ of a Lorentz transformation $L_0 = (b, I)$ shall contain all the transformations $L = (a, \Lambda)$ for which $| a_\kappa - b_\kappa | < \delta$ and $| \Lambda_{ik} - I_{ik} | < \delta$. The representation up to a factor $D(L)$ is continuous if there is to every positive number ϵ, every normalized wave function φ and every Lorentz transformation L_0 such a neighborhood δ of L_0 that for every L of this neighborhood one can find an Ω of modulus 1 (the Ω depending on L and φ) such that $(u_\varphi, u_\varphi) < \epsilon$ where

$$u_\varphi = (D(L_0) - \Omega D(L))\varphi. \tag{23}$$

Let us now take a point L_0 in the group space and find a normalized wave function φ for which $| (\varphi, D(L_0)\varphi) | > 1/6$. There always exists a φ with this property, if $|(\varphi, D(L_0)\varphi)| < 1/6$ then $\psi = \alpha\varphi + \beta D(L_0)\varphi$ with suitably chosen α and β will be normalized and $| (\psi, D(L_0)\psi) | > 1/6$. We consider then such a neighborhood $\mathfrak{N}$ of L_0 for all L of which $| (\varphi, D(L)\varphi) | > 1/12$. It is well known[19] that the whole group space can be covered with such neighborhoods. We want to show now that the $D(L)\varphi$ can be multiplied with such phase factors (depending on L) of modulus unity that it becomes strongly continuous in the region $\mathfrak{N}$.

We shall chose that phase factor so that $(\varphi, D(L)\varphi)$ becomes real and positive. Denoting then

$$(D(L_1) - D(L))\varphi = U_\varphi, \tag{23'}$$

the (U_φ, U_φ) can be made arbitrarily small by letting L approach sufficiently near to L_1, if L_1 is in $\mathfrak{N}$. Indeed, on account of the continuity, as defined above, there is an $\Omega = e^{i\kappa}$ such that $(u, u) < \epsilon$ if L is sufficiently near to L_1 where

$$u = (D(L_1) - e^{i\kappa}D(L))\varphi.$$

[19] This condition is the "separability" of the group. Cf. e.g. A. Haar, Ann. of Math., 34, 147, 1933.

Taking the absolute value of the scalar product of u with φ one obtains

$$| (\varphi, D(L_1)\varphi) - \cos\kappa(\varphi, D(L)\varphi) - i \sin\kappa(\varphi, D(L)\varphi) | = | (\varphi, u) | \leqq \sqrt{\epsilon},$$

because of Schwartz's inequality. If only $\sqrt{\epsilon} < 1/12$, the κ must be smaller than $\pi/2$ because the absolute value is certainly greater than the real part, and both $(\varphi, D(L_1)\varphi)$ and $(\varphi, D(L)\varphi)$ are real and greater than $1/12$.

As the absolute value is also greater than the imaginary part, we

$$\sin\kappa < 12\sqrt{\epsilon}.$$

On the other hand,

$$U_\varphi = u + (e^{i\kappa} - 1)D(L)\varphi,$$

and thus

$$(U_\varphi, U_\varphi)^{\frac{1}{2}} \leqq (u, u)^{\frac{1}{2}} + | e^{i\kappa} - 1 | \leqq \sqrt{\epsilon} + 2 \sin\kappa/2$$

$$(U_\varphi, U_\varphi) \leqq 625\ \epsilon.$$

(b) It shall be shown next that if $D(L)\varphi$ is strongly continuous in a region and $D(L)$ is continuous in the sense defined at the beginning of this section, then $D(L)\psi$ with an arbitrary ψ is (strongly) continuous in that region also. We shall see, hence, that the $D(L)$, with any normalization which makes a $D(L)\varphi$ strongly continuous, is continuous in the ordinary sense: There is to every L_1, ϵ and *every* ψ a δ so that $(U_\psi, U_\psi) < \epsilon$ where

$$U_\psi = (D(L_1) - D(L))\psi$$

if L is in the neighborhood δ of L_1.

It is sufficient to show the continuity of $D(L)\psi$ where ψ is orthogonal to φ. Indeed, every ψ' can be decomposed into two terms, $\psi' = \alpha\varphi + \beta\psi$ the one of which is parallel, the other perpendicular to φ. Since $D(L)\varphi$ is continuous, according to supposition, $D(L)\psi' = \alpha D(L)\varphi + \beta D(L)\psi$ will be continuous also if $D(L)\psi$ is continuous.

The continuity of the representation up to a factor requires that it is possible to achieve that $(u_\psi, u_\psi) < \epsilon$ and $(u_{\psi+\varphi}, u_{\psi+\varphi}) < \epsilon$ where

$$u_\psi = (D(L_1) - \Omega_\psi D(L))\psi, \tag{23a}$$

$$u_{\psi+\varphi} = (D(L_1) - \Omega_{\psi+\varphi}D(L))(\psi + \varphi), \tag{23b}$$

with suitably chosen Ω's. According to the foregoing, it also is possible to choose L and L_1 so close that $(U_\varphi, U_\varphi) < \epsilon$.

Subtracting (23′) and (23a) from (23b) and applying $D(L)^{-1}$ on both sides gives

$$(\Omega_\psi - \Omega_{\psi+\varphi})\psi + (1 - \Omega_{\psi+\varphi})\varphi = D(L)^{-1}(u_{\psi+\varphi} - u_\psi - U_\varphi)$$

The scalar product of the right side with itself is less than 9ϵ. Hence both $| \Omega_\psi - \Omega_{\psi+\varphi} | < 3\epsilon^{\frac{1}{2}}$ and $| 1 - \Omega_{\psi+\varphi} | < 3\epsilon^{\frac{1}{2}}$ or $| 1 - \Omega_\psi | < 6\epsilon^{\frac{1}{2}}$. Because of $U_\psi = u_\psi - (1 - \Omega_\psi)D(L)\psi$, the $(U_\psi, U_\psi)^{\frac{1}{2}} < (u_\psi, u_\psi)^{\frac{1}{2}} + | 1 - \Omega_\psi |$ and thus $(U_\psi, U_\psi) < 49\epsilon$.

This completes the proof of the theorem stated under (b). It also shows that not only the continuity of $D(L)\varphi$ has been achieved in the neighborhood of L_0 by the normalization used in (a) but also that of $D(L)\psi$ with every ψ, i.e., the continuity of $D(L)$.

It is clear also that every finite part of the group space can be covered by a finite number of neighborhoods in which $D(L)$ can be made continuous. It is easy to see that the ω of (22) will be also continuous in these neighborhoods so that is is possible to make them continuous, apart from regions of lower dimensionality than their variables have. In the following only the fact will be used that they can be made continuous in the neighborhood of any a, b and Λ.

B.

(a) We want to show next that all $T(a)$ commute. From (22B) we have

$$T(a)T(b)T(a)^{-1} = c(a, b)T(b) \tag{24}$$

where $c(a, b) = \omega(a, b)/\omega(b, a)$ and hence

$$c(a, b) = c(b, a)^{-1}. \tag{24a}$$

Transforming (24) with $T(a')$ one obtains

$$T(a')T(a)T(b)T(a)^{-1}T(a')^{-1} = c(a, b)T(a')T(b)T(a')^{-1}$$

or $$\omega(a', a)T(a' + a)T(b)\omega(a', a)^{-1}T(a' + a)^{-1} = c(a, b)c(a', b)T(b)$$

or

$$c(a, b)c(a', b) = c(a + a', b). \tag{25}$$

It follows[20] from (25) and the partial continuity of $c(a, b)$ that

$$c(a, b) = \exp\left(2\pi i \sum_{\kappa=1}^{4} a_\kappa f_\kappa(b)\right) \tag{26}$$

and, since this is equal to $c(b, a)^{-1} = \exp(-2\pi i \sum b_\kappa f_\kappa(a))$

$$\sum_{\kappa=1}^{4} (a_\kappa f_\kappa(b) + b_\kappa f_\kappa(a)) = n(a, b) \tag{27}$$

where $n(a, b)$ is an integer. Setting in (27) for b the vector $e^{(\lambda)}$ the λ component of which is 1, all the others zero and for $f_\kappa(e^{(\lambda)}) = -f_{\kappa\lambda}$

$$f_\lambda(a) = n(a, e^{(\lambda)}) + \sum_\kappa a_\kappa f_{\kappa\lambda},$$

and putting this back into (27) we obtain

$$\sum_{\kappa,\lambda=1}^{4} f_{\kappa\lambda}(a_\lambda b_\kappa + b_\lambda a_\kappa) + \sum_{\kappa=1}^{4} a_\kappa n(b, e^{(\kappa)}) + b_\kappa n(a, e^{(\kappa)}) = n(a, b). \tag{28}$$

---

[20] G. Hamel, Math. Ann. 60, 460, 1905, quoted from H. Hahn, *Theorie der reellen Funktionen*. Berlin 1921, pages 581–583.

Assuming for the components of a and b such values which are transcendental both with respect to each other and the $f_{\kappa\lambda}$ (which are fixed numbers), one sees that (28) cannot hold except if the coefficient of every one vanishes

$$f_{\kappa\lambda} + f_{\lambda\kappa} = 0; \qquad n(b, e^{(\kappa)}) = 0, \tag{29}$$

so that (26) becomes

$$c(a, b) = \exp\left(2\pi i \sum_{\kappa,\lambda=1}^{4} f_{\kappa\lambda} a_\lambda b_\kappa\right). \tag{30}$$

It is necessary now to consider the existence of an operator $d(\Lambda)$ satisfying (22C). Transforming this equation with the similar equation containing b instead of a

$$d(\Lambda)T(b)d(\Lambda)^{-1}d(\Lambda)T(a)d(\Lambda)^{-1}d(\Lambda)T(b)^{-1}d(\Lambda)^{-1}$$

$$= \omega(\Lambda, b)T(\Lambda b)\omega(\Lambda, a)T(\Lambda a)\omega(\Lambda, b)^{-1}T(\Lambda b)^{-1} = \omega(\Lambda, a)c(\Lambda b, \Lambda a)T(\Lambda a),$$

while the first line is clearly $d(\Lambda)c(b, a)T(a)d(\Lambda)^{-1} = \omega(\Lambda, a)c(b, a)T(\Lambda a)$ whence

$$c(b, a) = c(\Lambda b, \Lambda a) \tag{31}$$

holds for every Lorentz transformation Λ. Combined with (30) this gives

$$\sum_{\kappa\lambda}\left(f_{\kappa\lambda} a_\kappa b_\lambda - \sum_{\nu\mu} f_{\nu\mu}\Lambda_{\nu\kappa}\Lambda_{\mu\lambda} a_\kappa b_\lambda\right) = n'(a, b),$$

where $n'(a, b)$ is again an integer. As this equation holds for every a, b

$$f_{\kappa\lambda} = \sum_{\nu\mu} f_{\nu\mu}\Lambda_{\nu\kappa}\Lambda_{\mu\lambda}; \qquad f = \Lambda' f \Lambda$$

must hold also, for every Lorentz transformation. However, the only form invariant under all Lorentz transformations are multiples of the F of (10). Actually, because of (29), f must vanish and $c(a, b) = 1$, all the operators corresponding to translations commute

$$T(a)T(b) = T(b)T(a). \tag{32}$$

It is well to remember that it was necessary for obtaining this result to use the existence of $d(\Lambda)$ satisfying (22C).

(b) Equation (32) is clearly independent of the normalization of the $T(a)$. If we could fix the translation operators in four linearly independent directions $e^{(1)}, e^{(2)}, e^{(3)}, e^{(4)}$ so that for each of these directions

$$T(ae^{(\kappa)})T(be^{(\kappa)}) = T((a + b)e^{(\kappa)}) \tag{33}$$

be valid for every pair of numbers a, b, then the normalization

$$T(a_1e^{(1)} + a_2e^{(2)} + a_3e^{(3)} + a_4e^{(4)}) = T(a_1e^{(1)})T(a_2e^{(2)})T(a_3e^{(3)})T(a_4e^{(4)}) \tag{33a}$$

and (32) would ensure the general validity of

$$T(a)T(b) = T(a + b). \tag{34}$$

As the four linearly independent directions $e^{(1)}, \cdots, e^{(4)}$ we shall take four null vectors. If e is a null vector, there is, according to section 3, a homogeneous Lorentz transformation[21] Λ_e such that $\Lambda_e e = 2e$.

We normalize $T(e)$ so that

$$d(\Lambda_e)T(e)d(\Lambda_e)^{-1} = T(e)^2. \tag{35}$$

This is clearly independent of the normalization of $d(\Lambda_e)$. We further normalize for all (positive and negative) integers n

$$d(\Lambda_e)^n T(e)d(\Lambda_e)^{-n} = T(2^n e). \tag{35a}$$

It follows from this equation also that

$$T(2^n e)^2 = d(\Lambda_e)^n T(e)^2 d(\Lambda_e)^{-n} = d(\Lambda_e)^n d(\Lambda_e)T(e)d(\Lambda_e)^{-1}d(\Lambda_e)^{-n} = T(2^{n+1}e). \tag{36}$$

This allows us to normalize for every positive integer k

$$T(k\cdot 2^{-n}e) = T(2^{-n}e)^k \tag{35b}$$

in such a way that the normalization remains the same if we replace k by $2^m k$ and n by $n + m$. This ensures, together with (36), the validity of

$$\begin{aligned} T(\nu e)T(\mu e) &= T((\nu + \mu)e) \\ d(\Lambda_e)T(\nu e)d(\Lambda_e)^{-1} &= T(2\nu e) \end{aligned} \tag{36a}$$

for all dyadic fractions ν and μ.

It must be shown that if $\nu_1, \nu_2, \nu_3, \cdots$ is a sequence of dyadic fractions, converging to 0, $\lim T(\nu_i e) = 1$. From $T(a)\cdot T(0) = \omega(a, 0)T(a)$ it follows that $T(0)$ is a constant. According to the theorem of part (A)(b), the $T(\nu e)$, if multiplied by proper constants Ω_ν will converge to 1, i.e., by choosing an arbitrary φ, it is possible to make both $(1 - \Omega_\nu T(\nu e))\varphi = u$ and $(1 - \Omega_\nu T(\nu e))\cdot d(\Lambda_e)^{-1}\varphi = u'$ arbitrarily small, by making ν small. Applying $d(\Lambda_e)$ to the second expression, one obtains, for (36a), that $(1 - \Omega_\nu T(2\nu e))\varphi = d(\Lambda_e)u'$ is also small. On the other hand, applying $T(\nu e)$ to the first expression one sees that $(T(\nu e) - \Omega_\nu T(2\nu e))\varphi = T(\nu e)u$ approaches zero also. Hence, the difference of these two quantities $(1 - T(\nu e))\varphi$ goes to zero, i.e. $T(\nu_i e)\varphi$ converges to φ if $\nu_1, \nu_2, \nu_3, \cdots$ is a sequence of dyadic fractions approaching 0.

Now $\nu_1, \nu_2, \nu_3, \cdots$ be a sequence of dyadic fractions converging to an arbitrary number a. It will be shown then that $T(\nu_i e)$ converges to a multiple of $T(ae)$ and this multiple of $T(ae)$ will be the normalized $T(ae)$. Again, it follows from the continuity that there are such Ω_i that $\Omega_i T(\nu_i e)\varphi$ converges to $T(ae)\varphi$. The $\Omega_j^{-1}T(\nu_j e)^{-1}\Omega_i T(\nu_i e)\varphi$ will converge to φ, therefore, as both i and j tend to infinity. However, according to the previous paragraph, $T((\nu_i - \nu_j)e)\varphi$ tends to φ and thus $\Omega_j^{-1}\Omega_i$ tends to 1. It follows that Ω_i^{-1} converges to a definite number Ω. Hence $\Omega_i^{-1}\cdot\Omega_i T(\nu_i e)\varphi$ converges to $\Omega T(ae)\varphi$ which will be denoted, henceforth, by $T(ae)$. For the $T(ae)$, normalized in this way, (33) will hold,

[21] The index e denotes here the vector e for which $\Lambda_e e = 2e$; this Λ_e has no elementary divisor.

since if $\mu_1, \mu_2, \mu_3, \cdots$ are dyadic fractions converging to b, we obtain, with the help (36a)

$$T(ae)T(be)\varphi = \lim_{i,j=\infty} T((\nu_i + \mu_j)e)\varphi = T((a+b)e)\varphi.$$

This argument not only shows that it is possible to normalize the $T(ae^{(k)})$ and hence by (33a) the $T(a)$ so that (34) holds for them but, in addition to this, that these $T(a)$ will be continuous in the ordinary sense.

C.

It is clear that (34) will remain valid if one replaces $T(a)$ by $\exp(2\pi i\{a, c\})T(a)$ where c is an arbitrary vector. This remaining freedom in the normalization of $T(a)$ will be used to eliminate the $\omega(\Lambda, a)$ from (22C).

Transforming (22C) $d(\Lambda)T(a)d(\Lambda)^{-1} = \omega(\Lambda, a)T(\Lambda a)$ with $d(M)$ one obtains on the left side $\omega(M, \Lambda)d(M\Lambda)T(a)\omega(M, \Lambda)^{-1}\, d(M\Lambda)^{-1} = \omega(M\Lambda, a)T(M\Lambda a)$ while the right side becomes $\omega(\Lambda, a)\omega(M, \Lambda a)T(M\Lambda a)$. Hence

$$\omega(M\Lambda, a) = \omega(M, \Lambda a)\omega(\Lambda, a). \tag{37}$$

On the other hand, the product of two equations (22C) with the same Λ but with a and b respectively, instead of a yields with the help of (34)

$$\omega(\Lambda, a)\omega(\Lambda, b) = \omega(\Lambda(a+b)).$$

Hence

$$\omega(\Lambda, a) = \exp(2\pi i\{a, f(\Lambda)\}),$$

where $f(\Lambda)$ is a vector which can depend on Λ. Inserting this back into (37) one obtains

$$\{a, f(M\Lambda)\} = \{\Lambda a, f(M)\} + \{a, f(\Lambda)\} + n,$$

$$\{a, f(M\Lambda) - \Lambda^{-1}f(M) - f(\Lambda)\} = n,$$

where n is an integer which must vanish since it is a linear function of a. Hence

$$f(M\Lambda) = \Lambda^{-1}f(M) + f(\Lambda). \tag{38}$$

If we can show that the most general solution of the equation is

$$f(\Lambda) = (\Lambda^{-1} - 1)v_0, \tag{39}$$

where v_0 is a vector independent of Λ, the $\omega(\Lambda, a)$ will become $\omega(\Lambda, a) = \exp(2\pi i\{(\Lambda - 1)a, v_0\})$. Then $\omega(\Lambda, a)$ in (22C) will disappear if we replace $T(a)$ by $\exp(2\pi i\{a, v_0\})T(a)$.

The proof that (39) is a consequence of (38) is somewhat laborious. One can first consider the following homogeneous Lorentz transformations

$$X(\alpha_1, \gamma_1) = \begin{Vmatrix} C_1 & 0 & 0 & S_1 \\ 0 & c_1 & s_1 & 0 \\ 0 & -s_1 & c_1 & 0 \\ S_1 & 0 & 0 & C_1 \end{Vmatrix}; \quad Y(\alpha_2, \gamma_2) = \begin{Vmatrix} c_2 & 0 & -s_2 & 0 \\ 0 & C_2 & 0 & S_2 \\ s_2 & 0 & c_2 & 0 \\ 0 & S_2 & 0 & C_2 \end{Vmatrix}$$

(40)

$$Z(\alpha_3, \gamma_3) = \begin{Vmatrix} c_3 & s_3 & 0 & 0 \\ -s_3 & c_3 & 0 & 0 \\ 0 & 0 & C_3 & S_3 \\ 0 & 0 & S_3 & C_3 \end{Vmatrix}$$

where $c_i = \cos \alpha_i$; $s_i = \sin \alpha_i$; $C_i = \mathrm{Ch}\gamma_i$; $S_i = \mathrm{Sh}\gamma_i$. All the $X(\alpha, \gamma)$ commute. Let us choose, therefore, two angles α_1, γ_1 for which $1 - X(\alpha_1, \gamma_1)^{-1}$ has a reciprocal. It follows then from (38)

$$X(\alpha, \gamma)^{-1} f(X(\alpha_1, \gamma_1)) + f(X(\alpha, \gamma)) = X(\alpha_1, \gamma_1)^{-1} f(X(\alpha, \gamma)) + f(X(\alpha_1, \gamma_1))$$

(41) or $$f(X(\alpha, \gamma)) = [1 - X(\alpha_1, \gamma_1)^{-1}]^{-1}[1 - X(\alpha, \gamma)^{-1}] f(X(\alpha_1, \gamma_1))$$

$$f(X(\alpha, \gamma)) = (1 - X(\alpha, \gamma)^{-1}) v_X,$$

where v_X is independent of α, γ. Similar equations hold for the $f(Y(\alpha, \gamma))$ and $f(Z(\alpha, \gamma))$. Let us denote now $X(\pi, 0) = X$; $Y(\pi, 0) = Y$; $Z(\pi, 0) = Z$. These anticommute in the following sense with the transformations (40):

(42) $$YX(\alpha, \gamma)Y = ZX(\alpha, \gamma)Z = X(\alpha, \gamma)^{-1}.$$

From (38) one easily calculates

$$f(YX(\alpha, \gamma)Y) = (YX(\alpha, \gamma)^{-1} + 1) f(Y) + Yf(X(\alpha, \gamma)),$$

or, because of (41) and (42), after some trivial transformations

(43) $$(1 - X(\alpha, \gamma))(1 - Y)(v_X - v_Y) = 0.$$

As α, γ can be taken arbitrarily, the first factor can be dropped. This leaves $(1 - Y)(v_X - v_Y) = 0$, or that the first and third components of v_X and v_Y are equal. One similarly concludes, however, that $(1 - X)(v_Y - v_X) = 0$ and thus that the first three components of v_X, v_Y and also of v_Z are equal.

For $\gamma_1 = \gamma_2 = \gamma_3 = 0$ the transformations (40) are the generators of all rotations, i.e. all Lorentz transformations R not affecting the fourth coördinate. As the 4–4 matrix element of these transformations is 1, the expression $(1 - R^{-1})v$ is independent of the fourth component of v and $(1 - R^{-1})v_X = (1 - R^{-1})v_Y = (1 - R^{-1})v_Z$. It follows from (38) that if $f(R) = (1 - R^{-1})v_X$ and $f(S) = (1 - S^{-1})v_X$, then $f(SR) = (1 - R^{-1}S^{-1})v_X$. Thus $f(R) = (1 - R^{-1})v_X$ is valid with the same v_X for all rotations.

Now

$$f(X(\alpha, \gamma)R) = R^{-1}(1 - X(\alpha, \gamma)^{-1})v_X + (1 - R^{-1})v_X = (1 - (X(\alpha, \gamma)R)^{-1})v_X.$$

One easily concludes from (38) that the $f(E)$ corresponding to the unit operation vanishes and $f(\Lambda^{-1}) = -\Lambda f(\Lambda)$. Hence $f(R^{-1}X(\alpha, \gamma)^{-1}) = (1 - X(\alpha, \gamma)R)v_X$; and one concludes further that for all Lorentz transformations $\Lambda = RX(\alpha, \gamma)S$, (39) holds with $v_0 = -v_X$ if R and S are rotations. However, every homogeneous Lorentz transformation can be brought into this form (Section 4C). This completes the proof of (39) and thus of $\omega(\Lambda, a) = 1$.

D.

The quantities $\omega(a, b)$ and $\omega(\Lambda, a)$ for which it has just been shown that they can be assumed to be 1, are independent from the normalization of $d(\Lambda)$. We can affix therefore an arbitrary factor of modulus 1 to all the $d(\Lambda)$, without interfering with the normalizations so far accomplished. In consequence hereof, the ensuing discussion will be simply a discussion of the normalization

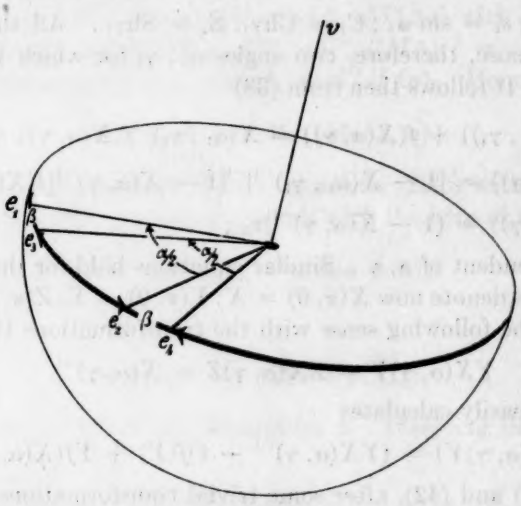


FIG. 2

of the operators for the homogeneous Lorentz group and the result to be obtained will be valid for that group also.

Partly because the representations up to a factor of the three dimensional rotation group may be interesting in themselves, but more particularly because the procedure to be followed for the Lorentz group can be especially simply demonstrated for this group, the three dimensional rotation group shall be taken up first.

It is well known that the normalization cannot be carried so far that $\omega(\Lambda, I) = 1$ in (22D) and there are well known representations for which $\omega(\Lambda, I) = \pm 1$. We shall allow this ambiguity therefore from the outset.

One can observe, first, that the operator corresponding to the unity of the group is a constant. This follows simply from $d(\Lambda)d(E) = \omega(\Lambda, E)d(\Lambda)$. The square of an operator corresponding to an involution is a constant, therefore.

The operator corresponding to the rotation about the axis e by the angle π, normalized so that its square be actually 1, will be denoted by $\tilde{e}$; $\tilde{e}^2 = 1$. The $\tilde{e}$ are—apart from the sign—uniquely defined.

A rotation R about v by the angle α is the product of two rotations by π about e_1 and e_2 where e_1 and e_3 are perpendicular to v and e_2 arises from e_1 by rotation about v with $\alpha/2$. Choosing for every v an arbitrary e_1 perpendicular to v, we can normalize, therefore

$$d(R) = \pm\tilde{e}_1\tilde{e}_2 . \tag{44}$$

Now $d(R)$ commutes with every $d(S)$ if S is also a rotation about v. This is proved in equations (24)–(30). The f_{11} in (30) must vanish on account of (29).

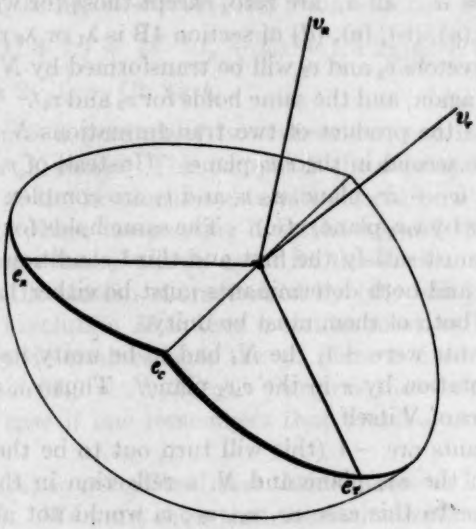


FIG. 3

(Also, both R and S can be arbitrarily accurately represented as powers of a very small rotation about v). Hence, transforming (44) by $d(S)$ one obtains

$$d(R) = \pm d(S)\tilde{e}_1 d(S)^{-1}\cdot d(S)\tilde{e}_2 d(S)^{-1}. \tag{44a}$$

Now $d(S)\tilde{e}_1 d(S)^{-1}$ corresponds to a rotation by π about an axis, perpendicular to v and enclosing an angle β with e_1, where β is the angle of rotation of S. Since the square of $d(S)\tilde{e}_1 d(S)^{-1}$ is also 1, (44a) is simply another way of writing $d(R) = \tilde{e}_3\tilde{e}_4$ as a product of two $\tilde{e}$ and we see that the normalization (44) is independent of the choice of the axis e_1 (Cf. Fig. 2).

For computing $d(R)d(T)$ we can draw the planes perpendicular to the axes of rotation of R and T and use for $d(R) = \tilde{e}_R\tilde{e}_C$ such a development that the axis e_C of the second involution coincide with the intersection line of the above-mentioned planes, while for $d(T) = \tilde{e}_C\tilde{e}_T$ we choose the first involution to be a rotation about this intersection line (Fig. 3). Then, the product

$$d(R)d(T) = \pm\tilde{e}_R\tilde{e}_C\tilde{e}_C\tilde{e}_T = \pm\tilde{e}_R\tilde{e}_T \tag{45}$$

will automatically have the normalization corresponding to (44). This shows that the operators normalized in (44) give a representation up to the sign.

For the Lorentz group, the proof can be performed along the same line, only the underlying geometrical facts are less obvious. Let Λ be a Lorentz transformation without elementary divisors with the characteristic values $e^{2i\gamma}$, $e^{-2i\gamma}$, $e^{2\chi}$, $e^{-2\chi}$ and the characteristic vectors v_1, $v_2 = v_1^*$, v_3, v_4, as described in section 4B.

We want to make $\Lambda = MN$ with $M^2 = N^2 = 1$. For $\Lambda N = M$, we have $\Lambda N \Lambda N = 1$ and thus $\Lambda N \Lambda = N$. Setting $Nv_i = \sum_k \alpha_{ik} v_k$, we obtain $\Lambda N \Lambda v_i = \sum \lambda_k \alpha_{ik} \lambda_i v_i = \sum \alpha_{ik} v_k$. Because of the linear independence of the v_k this amounts to $\lambda_i \lambda_k \alpha_{ik} = \alpha_{ik}$: all α_{ik} are zero, except those for which $\lambda_i \lambda_k = 1$. As in none of the cases (a), (b), (c), (d) of section 4B is λ_1 or λ_2 reciprocal to one of the last two λ, the vectors v_1 and v_2 will be transformed by N into a linear combination of v_1 and v_2 again, and the same holds for v_3 and v_4. This means that N can be considered as the product of two transformations $N = N_s N_t$, the first in the $v_1 v_2$ plane, the second in the $v_3 v_4$ plane. (Instead of $v_1 v_2$ plane one really should say $v_1 + v_2$, $iv_1 - iv_2$ plane, as v_1 and v_2 are complex themselves. This will be meant always by $v_1 v_2$ plane, etc.). The same holds for M also.

Both N_s and N_t must satisfy the first and third condition for Lorentz transformations (cf. 4A) and both determinants must be either 1, or -1. Furthermore, the square of both of them must be unity.

If both determinants were $+1$, the N_t had to be unity itself, while N_s could be the unity or a rotation by π in the $v_1 v_2$ plane. Thus v_1, v_2, v_3, v_4 would be characteristic vectors of N itself.

If both determinants are -1 (this will turn out to be the case), N_s is a reflection on a line in the $v_1 v_2$ plane and N_t a reflection in the $v_3 v_4$ plane, interchanging v_3 and v_4. In this case v_1, v_2, v_3, v_4 would not all be characteristic vectors of N.

If v_1, v_2, v_3, v_4 are characteristic vectors of N, they are characteristic vectors of $M = \Lambda N$ also. Then both M and N would be either unity, or a rotation by π in the $v_1 v_2$ plane. If both of them were rotations in the $v_1 v_2$ plane, their product Λ would be the unity which we want to exclude for the present. We can exclude the remaining cases in which the determinants of N_s and N_t are $+1$ by further stipulating that neither M nor N shall be the unity in the decomposition $\Lambda = MN$.

Hence N is the product of a reflection in the $v_1 v_2$ plane

$$Ns_\nu' = s_\nu' ; \qquad Ns_\nu = -s_\nu \tag{46a}$$

where s_ν and s_ν' are two perpendicular real vectors in the $v_1 v_2$ plane

$$s_\nu' = e^{i\nu} v_1 + e^{-i\nu} v_2 ; \qquad s_\nu = i(e^{i\nu} v_1 - e^{-i\nu} v_2), \tag{46b}$$

and of a reflection in the $v_3 v_4$ plane

$$Nt_\mu' = t_\mu' ; \qquad Nt_\mu = -t_\mu , \tag{46c}$$

where again t_μ , t'_μ are real vectors in the v_3v_4 plane, perpendicular to each other, t_μ being space-like, t'_μ time-like:

$$t'_\mu = e^{\mu}v_3 + e^{-\mu}v_4 ; \qquad t_\mu = e^{\mu}v_3 - e^{-\mu}v_4 . \tag{46d}$$

Thus N becomes a rotation by π in the purely space like s,t_μ plane. The M can be calculated from $M = \Lambda N$

$$\begin{aligned} Ms'_\nu &= \Lambda N s'_\nu = \Lambda s'_\nu = e^{i\nu+2i\gamma}v_1 + e^{-i\nu-2i\gamma}v_2 \\ &= \tfrac{1}{2}e^{2i\gamma}(s'_\nu - is_\nu) + \tfrac{1}{2}e^{-2i\gamma}(s'_\nu + is_\nu) = \cos 2\gamma\cdot s'_\nu + \sin 2\gamma\cdot s_\nu \\ Ms_\nu &= \sin 2\gamma\cdot s'_\nu - \cos 2\gamma\cdot s_\nu \\ Mt'_\mu &= \Lambda N t'_\mu = \Lambda t'_\mu = e^{\mu+2\chi}v_3 + e^{-\mu-2\chi}v_4 \\ &= \tfrac{1}{2}e^{2\chi}(t'_\mu + t_\mu) + \tfrac{1}{2}e^{-2\chi}(t'_\mu - t_\mu) = \mathrm{Ch}\ 2\chi\cdot t'_\mu + \mathrm{Sh}\ 2\chi\cdot t_\mu \\ Mt_\mu &= -\mathrm{Sh}\ 2\chi\cdot t'_\mu - \mathrm{Ch}\ 2\chi\cdot t_\mu . \end{aligned} \tag{46e}$$

Thus M also becomes a product of two reflections, one in the $v_1v_2 = s'_\nu s_\nu$ the other in the $v_3v_4 = t'_\mu t_\mu$ plane. This completes the decomposition of Λ into two involutions. One of the involutions can be taken to be a rotation by π in an arbitrary space like plane, intersecting both the v_1v_2 and the v_3v_4 planes, as the freedom in choosing ν and μ allows us to fix the lines s_ν and t_μ arbitrarily in those planes. The involution characterized by (46) will be called $N_{\nu\mu}$ henceforth. The other involution M is then a similar rotation, in a plane, however, which is completely determined once the s,t_μ plane is fixed. It will be denoted by $M_{\nu\mu}$ (it is, in fact $M_{\nu\mu} = N_{\nu+\gamma\ \mu+\chi}$). One sees the complete analogy to the three dimensional case if one remembers that γ and χ are the half angles of rotation.

The $d(M)$ and $d(N)$ so normalized that their squares be 1 shall be denoted by $d_1(M_{\nu\mu})$ and $d_1(N_{\nu\mu})$. We must show that the normalization for

$$d(\Lambda) = \pm d_1(M_{\nu\mu})d_1(N_{\nu\mu}) \tag{47}$$

is independent of ν and μ. For this purpose, we transform

$$d(\Lambda) = \pm d_1(M_{00})d_1(N_{00}) \tag{47a}$$

with $d(\Lambda_1)$ where Λ_1 has the same characteristic vectors as Λ but different characteristic values, namely $e^{i\nu}$, $e^{-i\nu}$, e^{μ} and $e^{-\mu}$. Since $\Lambda_1 M_{00}\Lambda_1^{-1} = M_{\nu\mu}$ and $\Lambda_1 N_{00}\Lambda_1^{-1} = N_{\nu\mu}$ we have $d(\Lambda_1)d_1(M_{00})d(\Lambda_1)^{-1} = \omega d_1(M_{\nu\mu})$ where $\omega = \pm 1$, as the squares of both sides are 1. Hence, (47a) becomes if transformed with $d(\Lambda_1)$ just

$$d(\Lambda_1)d(\Lambda)d(\Lambda_1)^{-1} = \pm d_1(M_{\nu\mu})d_1(N_{\nu\mu}). \tag{47b}$$

The normalization (47) would be clearly independent of ν and μ if $d(\Lambda_1)$ commuted with $d(\Lambda)$.

Again, the argument contained in equations (24)-(30) can be applied and shows that

$$d(\Lambda_1)d(\Lambda)d(\Lambda_1)^{-1} = \exp\,(2\pi i f(2\gamma\mu - 2\chi\nu))d(\Lambda) \tag{48}$$

holds for every γ, χ, ν, μ. However, the exponential in (48) must be 1 if $\gamma = 0$; $\nu = 2\pi/n$; $\chi = \frac{1}{2}n\mu$ since in this case $\Lambda = \Lambda_1^n$. Thus $\exp(-4\pi^2 if\mu) = 1$ for every μ and $f = 0$ and the left side of (47b) can be replaced by $d(\Lambda)$; the normalization in (47) is independent of ν and μ.

In order to have the analogue of (45), we must show that, having two Lorentz transformations $\Lambda = M_{\nu\mu}N_{\nu\mu}$ and $I = P_{\alpha\beta}Q_{\alpha\beta}$ we can chose ν, μ and α, β so that $N_{\nu\mu} = P_{\alpha\beta}$ i.e. that the plane of rotation $s_\nu t_\mu$ of $N_{\nu\mu}$ coincide with the plane of rotation of $P_{\alpha\beta}$. As the latter plane can be made to an arbitrary space-like plane intersecting both the w_1w_2 and the w_3w_4 planes (where w_1, w_2, w_3, w_4 are the characteristic vectors of I), we must show the existence of a space like plane, intersecting all four planes v_1v_2, v_3v_4, w_1w_2, w_3w_4. Both the first and the second pair of planes are orthogonal.

One can show[22] that if Λ and I have no common null vector as characteristic

[22] We first suppose the existence of a real plane p intersecting all four planes v_1v_2, v_3v_4, w_1w_2, w_3w_4. If p intersects v_1v_2 the plane q perpendicular to p will intersect the plane v_3v_4 perpendicular to v_1v_2. Indeed, the line which is perpendicular to both p and v_1v_2 (there is such a line as p and v_1v_2 intersect) is contained in both q and v_3v_4. This shows that if there is a plane intersecting all four planes, the plane perpendicular to this will have this property also.

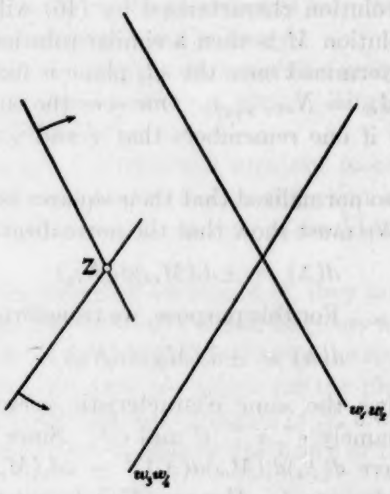


Fig. 4 gives a projection of all lines into the x_1x_2 plane. One sees that there are, in general, two intersecting planes, only in exceptional cases is there only one.

If the plane p—the existence of which we suppose for the time being—contains a time-like vector, q will be space-like (Section 4B, [1]). Both in this case and if p contains only space-like vectors, the theorem in the text is valid. There is a last possibility, that p is tangent to the light cone, i.e. contains only space like vectors and a null vector v. The space-like vectors of p are all orthogonal to v, otherwise p would contain time-like vectors also. In this case the plane q, perpendicular to p will contain v also. The line in which v_1v_2 intersects p is space-like and orthogonal to the vector in which v_3v_4 intersects p. The latter intersection must coincide with v, therefore, as no other vector of p is orthogonal to

vector, there are always two planes, perpendicular to each other which intersect four such planes. One of these is always space like. It is possible to assume, therefore, that both $N_{\nu\mu}$ and $P_{\alpha\beta}$ are the rotation by π in this plane. Thus

$$d(\Lambda)d(I) = \pm d_1(M_{\nu\mu})d_1(N_{\nu\mu})d_1(P_{\alpha\beta})d_1(Q_{\alpha\beta}) = \pm d_1(M_{\nu\mu})d_1(Q_{\alpha\beta}), \tag{49}$$

and $d(\Lambda)d(I)$ has the normalization corresponding to the product of two involutions, neither of which is unity. This is, however, also the normalization adopted for $d(\Lambda I)$. Hence

$$d(\Lambda)d(I) = \pm d(\Lambda I) \tag{49a}$$

holds if Λ, I and ΛI are Lorentz transformations corresponding to one of the cases (a), (b), (c) or (d) of section 4B and if Λ and I have no common characteristic null vector. In addition to this (49a) holds also, assuming $d(E) = \pm 1$, if any of the transformations Λ, I, ΛI is unity, or if both characteristic null vectors of Λ and I are equal, as in this case the planes v_3v_4 and w_3w_4 and also v_1v_2 and w_1w_2 coincide and there are many space like planes intersecting all.

If Λ and I have one common characteristic null vector, $v_3 = w_3$, the others, v_4 and w_4 respectively, being different, one can use an aritifice to prove (49a) which will be used in later parts of this section extensively. One can find a Lorentz transformation J so that none of the pairs $I - J$; $\Lambda - IJ$; $\Lambda IJ - J^{-1}$ has a common characteristic null vector. This will be true, e.g. if the characteristic null vectors of J are v_4 and another null vector, different from v_3, w_4 and the characteristic vectors of ΛI. Then (49a) will hold for all the above pairs and

$$d(\Lambda)d(I) = \pm d(\Lambda)d(I)d(J)d(J^{-1}) = \pm d(\Lambda)d(IJ)d(J^{-1}) = \pm d(\Lambda IJ)d(J^{-1}) = \pm d(\Lambda I).$$

---

any space-like vector in it. Hence, v is the intersection of p and v_3v_4 and is either v_3 or v_4. One can conclude in the same way that v coincides with either w_3 or w_4 also and we see that if p is tangent to the light cone the two transformations Λ and I have a common null vector as characteristic vector. Thus the theorem in the text is correct if we can show the existence of an arbitrary real plane p intersecting all four planes v_1v_2, v_3v_4, w_1w_2, w_3w_4.

Let us draw a coördinate system in our four dimensional space, the x_1x_2 plane of which is the v_1v_2 plane, the x_3 and x_4 axes having the directions of the vectors $v_3 - v_4$ and $v_3 + v_4$, respectively. The three dimensional manifold M characterized by $x_4 = 1$ intersects all planes in a line, the v_1v_2 plane in the line at infinity of the x_1x_2 plane, the v_3v_4 plane in the x_3 axis. The intersection of M with the w_1w_2 and w_3w_4 planes will be lines in M with directions perpendicular to each other. They will have a common normal through the origin of M, intersecting it at reciprocal distances. This follows from their orthogonality in the four dimensional space.

A plane intersecting v_1v_2 and v_3v_4 will be a line parallel to x_1x_2 through the x_3 axis. If we draw such lines through all points of the line corresponding to w_1w_2, the direction of this line will turn by π if we go from one end of this line to the other. Similarly, the lines going through the line corresponding to w_3w_4 will turn by π in the *opposite* direction. Thus the first set of lines will have at least one line in common with the second set and this line will correspond to a real plane intersecting all four planes v_1v_2, v_3v_4, w_1w_2, w_3w_4. This completes the proof of the theorem referred to in the text.

This completes the proof of (49a) for all cases in which Λ, I and ΛI have no elementary divisors. It is evident also that we can replace in the normalization (47) the d_1 by d. One also concludes easily that $d(M)^2$ is in the same representation either $+1$ for all involutions M, or -1 for every involution. The former ones will give real representations, the latter ones representations up to the sign.

If Λ has an elementary divisor, it can be expressed in the v_e, w_e, z_e, v_1 scheme as the matrix (Cf. equ. (20))

$$\Lambda_e = \begin{Vmatrix} 1 & 1 & \frac{1}{2} & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{Vmatrix}$$

and can be written, in the same scheme, as the product of two Lorentz transformations with the square 1

$$\Lambda_e = M_0 N_0 = \begin{Vmatrix} 1 & -1 & \frac{1}{2} & 0 \\ 0 & -1 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 \end{Vmatrix} \cdot \begin{Vmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 \end{Vmatrix}.$$

We can normalize therefore $d(\Lambda_e) = \pm d(M_0)d(N_0)$. If Λ can be written as the product of two other involutions also $\Lambda_e = M_1 N_1$ the corresponding normalization will be identical with the original one. In order to prove this, let us consider a Lorentz transformation J such that neither of the Lorentz transformations J, $N_0 J$, $N_1 J$, $\Lambda_e J = M_0 N_0 J = M_1 N_1 J$ have an elementary divisor. Since the number of free parameters is only 4 in case (e), while 6 for case (a), this is always possible. Then, for (45a)

$$\begin{aligned} d(M_0)d(N_0)d(J) &= \pm d(M_0)d(N_0 J) = \pm d(M_0 N_0 J) \\ &= \pm d(M_1 N_1 J) = \pm d(M_1)d(N_1 J) = \pm d(M_1)d(N_1)d(J) \end{aligned}$$

and thus $d(M_0)d(N_0) = \pm d(M_1)d(N_1)$. This shows also that even if ΛI is in case (e), $\omega(\Lambda, I) = \pm 1$, since (49) leads to the correct normalization.

If $\Lambda = MN$ has an elementary divisor, I not, $d(\Lambda)d(I)$ still will have the normalization corresponding to the product of two involutions. One can find again a J such that neither of the transformations J, J^{-1}, IJ, NIJ, $MNIJ$, have an elementary divisor. Then

$$\begin{aligned} d(M)d(N)d(I) &= \pm d(M)d(N)d(I)d(J)d(J)^{-1} \\ &= \pm d(M)d(N)d(IJ)d(J)^{-1} = \pm d(M)d(NIJ)d(J)^{-1} \\ &= d(\Lambda IJ)d(J^{-1}). \end{aligned}$$

The last product has, however, the normalization corresponding to two involutions, as was shown in (49a), since neither ΛIJ, nor J^{-1} is in case (e).

Lastly, we must consider the case when both Λ and I may have an elementary divisor. In this case, we need a J such that neither of J, J^{-1}, IJ have one. Then, because of the generalization of (49a) just proved, in which the first factor is in case (e)

$$d(\Lambda)d(I) = \pm d(\Lambda)d(I)d(J)d(J^{-1}) = \pm d(\Lambda)d(IJ)d(J^{-1})$$
$$= \pm d(\Lambda IJ)d(J^{-1})$$

which has the right normalization.

This completes the proof of

$$\omega(\Lambda, I) = \pm 1 \tag{50}$$

for all possible cases, and the normalization of all $D(L)$ of a representation of the inhomogeneous Lorentz group up to a factor, is carried out in such a way that the normalized operators give a representation up to the sign. It is even carried so far that in the first two of equations (22) $\omega = 1$ can be set. We shall consider henceforth systems of operators satisfying (7), or, more specifically, (22B) and (22C) with $\omega(a, b) = \omega(\Lambda, a) = 1$ and (22D) with $\omega(\Lambda, I) = \pm 1$.

E.

Lastly, it shall be shown that the renormalization not only did not spoil the partly continuous character of the representation, attained at the first normalization in part (A) of this section, but that the same holds now *everywhere*, in the ordinary sense for $T(a)$ and, apart from the ambiguity of sign, also for $d(\Lambda)$. For $T(a)$ this was proved in part $(B)(b)$ of this section, for $d(\Lambda)$ it means that to every Λ_1, ϵ and φ there is such a δ that *one* of the two quantities

$$((d(\Lambda_1) \mp d(\Lambda))\varphi, (d(\Lambda_1) \mp d(\Lambda))\varphi) < \epsilon \tag{51}$$

if Λ is in the neighborhood δ of Λ_1. The inequality (51) is equivalent to

$$((1 \mp d(\Lambda_0))\varphi, (1 \mp d(\Lambda_0))\varphi) < \epsilon, \tag{51a}$$

where $\Lambda_0 = \Lambda_1^{-1}\Lambda$ now can be assumed to be in the neighborhood of the unity. Thus, the continuity of $d(\Lambda)$ at $\Lambda = E$ entails the continuity everywhere.[23] In fact, it would be sufficient to show that the $d(X)$, $d(Y)$ and $d(Z)$ corresponding to the transformations (40) converge to ± 1, as α, γ approach 0, since one can write every transformation in the neighborhood of the unit element as a product $\Lambda = Z(0, \gamma_3)Y(0, \gamma_2)X(0, \gamma_1)X(\alpha_1, 0)Y(\alpha_2, 0)Z(\alpha_3, 0)$ and the parameters $\alpha_1, \cdots, \gamma_3$ will converge to 0 as Λ converges to 1. However, we shall carry out the proof for an arbitrary Λ without an elementary divisor.

For $d(\Lambda)$, equations (46) show that as Λ approaches E (i.e., as γ and χ approach zero) both M_{00} and N_{00} approach the same involution, which we shall call K. Let us now consider a wave function $\psi = \varphi + d_1(K)\varphi$ or, if this vanishes $\psi = \varphi - d_1(K)\varphi$. We have $d_1(K)\psi = \pm\psi$. If Λ is sufficiently near to unity,

[23] J. von Neumann, Sitz. d. kön. Preuss. Akad. p. 76, 1927.

$d_1(N_{00})\psi$ will be sufficiently near to $\Omega d_1(K)\psi = \pm\Omega\psi$ and all we have to show is that Ω approaches ± 1. The same thing will hold for $d_1(M_{00})$. Indeed from $d_1(N_{00})\psi - \Omega\psi = u$ it follows by applying $d_1(N_{00})$ on both sides $\psi - \Omega^2\psi = (d_1(N_{00}) + \Omega)u$. As (u, u) goes to zero, Ω must go to ± 1, and consequently, also $d_1(N_{00})\psi$ goes to ψ or to $-\psi$. Applying $d_1(M_{00})$ to this, one sees that $d_1(M_{00})d_1(N_{00})\psi = d(\Lambda)\psi$ goes to $\pm\psi$ as Λ goes to unity. The argument given in (A)(b) shows that this holds not only for ψ but for every other function also, i.e. $d(\Lambda)$ converges to $\pm 1 = d(E)$ as Λ approaches E. Thus $d(\Lambda)$ is continuous in the neighborhood of E and hence everywhere.

According to the last remark in part 4, the operators $\pm d(\Lambda)$ form a single valued representation of the group of complex unimodular two dimensional matrices C. Let us denote the homogeneous Lorentz transformation which corresponds in the isomorphism to C by $\bar{C}$. Our task of solving the equs. (22) has been reduced to finding all single valued unitary representation of the group with the elements $[a, C] = [a, 1]\,[0, C]$, the multiplication rule of which is $[a, C_1]\,[b, C_2] = [a + \bar{C}_1 b, C_1 C_2]$. For the representations of this group $D[a, C] = T(a)d[C]$ we had

$$T(a)T(b) = T(a + b)$$

(52a)
$$d[C]T(a) = T(\bar{C}a)\,d[C]$$

$$d[C_1]\,d[C_2] = d[C_1 C_2].$$

It would be more natural, perhaps, from the mathematical point of view, to use henceforth this new notation for the representations and let the d depend on the C rather than on the $\bar{C}$ or Λ. However, in order to be reminded on the geometrical significance of the group elements, it appeared to me to be better to keep the old notation. Instead of the equations (22B), (22C), (22D) we have, then

(52B)
$$T(a)T(b) = T(a + b)$$

(52C)
$$d(\Lambda)T(a) = T(\Lambda a)\,d(\Lambda)$$

(52D)
$$d(\Lambda)\,d(I) = \pm d(\Lambda I).$$

6. Reduction of the Representations of the Inhomogeneous Lorentz Group to Representations of a "Little Group"

This section, unlike the other ones, will often make use of methods, which though commonly accepted in physics, must be further justified from a rigorous mathematical point of view. This has been done, in the meanwhile, by J. von Neumann in an as yet unpublished article and I am much indebted to him for his coöperation in this respect and for his readiness in communicating his results to me. A reference to his paper[24] will be made whenever his work is necessary for making inexact considerations of this section rigorous.

[24] J. von Neumann, Ann. of Math. to appear shortly.

A.

Since the translation operators all commute, it is possible[24] to introduce such a coördinate system in Hilbert space that the wave functions $\varphi(p, \zeta)$ contain momentum variables p_1, p_2, p_3, p_4 and a discrete variable ζ so that

$$T(a)\varphi(p, \zeta) = e^{i\{p,a\}}\varphi(p, \zeta). \tag{53}$$

p will stand for the four variables p_1, p_2, p_3, p_4.

Of course, the fact that the Lorentzian scalar product enters in the exponent, rather than the ordinary, is entirely arbitrary and could be changed by changing the signs of p_1, p_2, p_3.

The unitary scalar product of two wave functions is not yet completely defined by the requirements so far made on the coördinate system. It can be a summation over ζ and an arbitrary Stieltjes integral over the components of p:

$$(\psi, \varphi) = \sum_\zeta \int \psi(p, \zeta)^* \varphi(p, \zeta)\, df(p, \zeta). \tag{54}$$

The importance of introducing a weight factor, depending on p, for the scalar product lies not so much in the possibility of giving finite but different weights to different regions in p space. Such a weight distribution $g(p, \zeta)$ always could be absorbed into the wave functions, replacing all $\varphi(p, \zeta)$ by $\sqrt{g(p, \zeta)} \cdot \varphi(p, \zeta)$. The necessity of introducing the $f(p, \zeta)$ lies rather in the possibility of some regions of p having zero weight while, on the other hand, at other places points may have finite weights. On account of the definite metric in Hilbert space, the integral $\int df(p, \zeta)$ over any region r, for any ζ, is either positive, or zero, since it is the scalar product of that function with itself, which is 1 in the region r of p and the value ζ of the discrete variable, zero otherwise.

Let us now define the operators

$$P(\Lambda)\varphi(p, \zeta) = \varphi(\Lambda^{-1}p, \zeta). \tag{55}$$

This equation defines the function $P(\Lambda)\varphi$, which is, at the point p, ζ, as great as the function φ at the point $\Lambda^{-1}p$, ζ. The operator $P(\Lambda)$ is not necessarily unitary, on account of the weight factor in (54). We can easily calculate

$$P(\Lambda)T(a)\varphi(p, \zeta) = T(a)\varphi(\Lambda^{-1}p, \zeta) = e^{i\{\Lambda^{-1}p,a\}}\varphi(\Lambda^{-1}p, \zeta),$$

$$T(\Lambda a)P(\Lambda)\varphi(p, \zeta) = e^{i\{p,\Lambda a\}}P(\Lambda)\varphi(p, \zeta) = e^{i\{p,\Lambda a\}}\varphi(\Lambda^{-1}p, \zeta),$$

so that, for $\{\Lambda^{-1}p, a\} = \{p, \Lambda a\}$, we have

$$P(\Lambda)T(a) = T(\Lambda a)P(\Lambda). \tag{56}$$

This, together with (52C), shows that $d(\Lambda)P(\Lambda)^{-1} = Q(\Lambda)$ commutes with all $T(a)$ and, therefore, with the multiplication with every function of p, since the exponentials form a complete set of functions of p_1, p_2, p_3, p_4. Thus

$$d(\Lambda) = Q(\Lambda)P(\Lambda), \tag{57}$$

where $Q(\Lambda)$ is an operator in the space of the ζ alone[24] which can depend, however, on the particular value of p in the underlying space:

$$Q(\Lambda)\varphi(p, \zeta) = \sum_{\eta} Q(p, \Lambda)_{\zeta\eta}\varphi(p, \eta). \tag{57a}$$

Here, $Q(p, \Lambda)_{\zeta\eta}$ are the components of an ordinary (finite or infinite) matrix, depending on p and Λ. From (57), we obtain

$$\begin{aligned} d(\Lambda)\varphi(p, \zeta) &= \sum_{\eta} Q(p, \Lambda)_{\zeta\eta}P(\Lambda)\varphi(p, \eta) \\ &= \sum_{\eta} Q(p, \Lambda)_{\zeta\eta}\varphi(\Lambda^{-1}p, \eta). \end{aligned} \tag{57b}$$

As the exponentials form a complete set of functions, we can approximate the operation of multiplication with any function of p_1, p_2, p_3, p_4 by a linear combination

$$f(p)\varphi = \sum_{n} c_n T(a_n)\varphi. \tag{58}$$

If we choose $f(p)$ to be such a function that

$$f(p) = f(\Lambda p) \tag{58a}$$

the operation of multiplication with $f(p)$ will commute with all operations of the group. It commutes evidently with the $T(a)$ and the $Q(p, \Lambda)$, and on account of (56) and (58), (58a) also with $P(\Lambda)$. Thus the operation of (58) belongs to the centrum of the algebra of our representation. Since, however, we assume that the representation is factorial (cf. 2), the centrum contains only multiples of the unity and

$$f(p)\varphi(p, \zeta) = c\varphi(p, \zeta). \tag{58b}$$

This can be true only if φ is different from zero only for such momenta p which can be obtained from each other by homogeneous Lorentz transformations, because $f(p)$ needs to be equal to $f(p')$ only if there is a Λ which brings them into each other.

It will be sufficient, henceforth, to consider only such representations, the wave functions of which vanish except for such momenta which can be obtained from one by homogeneous Lorentz transformations. One can restrict, then, the definition domain of the φ to these momenta.

These representations can now naturally be divided into the four classes enumerated in section 3, and two classes contain two subclasses. There will be representations, the wave functions of which are defined for such p that

(1) $\{p, p\} = P > 0$ (3) $p = 0$

(2) $\{p, p\} = P = 0; p \neq 0$ (4) $\{p, p\} = P < 0.$

The classes 1 and 2 contain two sub-classes each. In the positive subclasses P_+ and 0_+ the time components of all momenta are $p_4 > 0$, in the negative

subclasses P_- and 0_- the fourth components of the momenta are negative. Class 3 will be denoted by 0_0. If P is negative, it has no index.

From the condition that $d(\Lambda)$ shall be a unitary operator, it is possible to infer[24] that one can introduce a coördinate system in Hilbert space in such a way that

$$\int_r df(p, \zeta) = \int_{\Lambda r} df(p, \eta) \tag{59}$$

if $Q(p, \Lambda)_{\zeta\eta} \neq 0$ for the p of the domain r. Otherwise, r is an arbitrary domain in the space of p_1, p_2, p_3, p_4 and Λr is the domain which contains Λp if r contains p. Equation (59) holds for all ζ, η, except for such pairs for which $Q(p, \Lambda)_{\zeta\eta} = 0$. It is possible, hence, to decompose the original representation in such a way that (59) holds within every reduced part. Neither $T(a)$ nor $d(\Lambda)$ can have matrix elements between such η and ζ for which (59) does not hold.

In the third class of representations, the variable p can be dropped entirely, and $T(a)\varphi(\zeta) = \varphi(\zeta)$, i.e., all wave functions are invariant under the operations of the invariant subgroup, formed by the translations. The equation $T(a)\varphi(\zeta) = \varphi(\zeta)$ is an invariant characterization of the representations of the third class, i.e., a characterization which is not affected by a similarity transformation. Hence, the reduced parts of a representation of class 3 also belong to this class.

Since no wave function of the other classes can remain invariant under all translations, no representation of the third class can be contained in any representation of one of the other classes. In the other classes, the variability domain of p remains three dimensional. It is possible, therefore, to introduce instead of p_1, p_2, p_3, p_4 three independent variables. In the cases 1 and 2 with which we shall be concerned most, p_1, p_2, p_3 can be kept for these three variables. On account of (59), the Stieltjes integral can be replaced by an ordinary integral[24] over these variables, the weight factor being $|p_4|^{-1} = (P + p_1^2 + p_2^2 + p_3^2)^{-\frac{1}{2}}$

$$\{\psi, \varphi\} = \sum_\zeta \int\int\int_{-\infty}^{\infty} \psi(p, \zeta)^* \varphi(p, \zeta) \, |p_4|^{-1} dp_1 dp_2 dp_3. \tag{59a}$$

In fact, with the weight factor $|p_4|^{-1}$ the weight of the domain r i.e., $W_r = \int\int\int_r |p_4|^{-1} dp_1 dp_2 dp_3$ is equal to the weight of the domain $W_{\Lambda r}$ as required[25] by (59). Having the scalar product fixed in this way, $P(\Lambda)$ becomes a unitary operator and, hence, $Q(\Lambda)$ will be unitary also.

We want to give next a characterization of the representations with a given P, which is independent of the coördinate system in Hilbert space. It follows from

---

[25] The invariance of integrals of the character of (59a) is frequently made use of in relativity theory. One can prove it by calculating the Jacobian of the transformation

$$p_i' = \Lambda_{i1} p_1 + \Lambda_{i2} p_2 + \Lambda_{i3} p_3 + (P + p_1^2 + p_2^2 + p_3^2)^{\frac{1}{2}} \qquad (i = 1, 2, 3)$$

which comes out to be $(P + p_1^2 + p_2^2 + p_3^2)^{\frac{1}{2}}(P + p_1'^2 + p_2'^2 + p_3'^2)^{-\frac{1}{2}}$. Equ. (59a) will not be used in later parts of this paper.

(53), that in a representation with a given P the wave functions $\psi_1, \psi_2, \cdots$ which are different from zero only in a finite domain of p, form an everywhere dense set, to all elements of which the infinitesimal operators of translation can be applied arbitrarily often

$$\lim_{h=0} h^{-n}(T(he) - 1)^n\psi = \lim_{h=0} h^{-n}(e^{ih\{p,e\}} - 1)^n\psi = i^n\{p, e\}^n\psi, \tag{60}$$

where e will be a unit vector in the direction of a coördinate axis or oppositely directed to it. Hence for all members ψ of this everywhere dense set

$$\lim_{h=0} \sum_k \pm h^{-2}(T(2he_k) - 2T(he_k) + 1)\psi = (p_1^2 + p_2^2 + p_3^2 - p_4^2)\psi = -P\psi, \tag{61}$$

where e_k is a unit vector in (or opposite) the k^{th} coördinate axis and the $\pm$ is $+$ for $k = 4$, and $-$ for $k = 1, 2, 3$.

On the other hand, there is no φ for which

$$\lim_{h=0} \sum_k \pm h^{-2}(T(2he_k) - 2T(he_k) + 1)\varphi \tag{61a}$$

if it exists, would be different from $-P\varphi$. Suppose the limit in (61a) exists and is $-P\varphi + \varphi'$. Let us choose then a normalized ψ, from the above set, such that $(\psi, \varphi') = \delta$ with $\delta > 0$ and an h so that the expression after the lim sign in (61a) assumes the value $-P\varphi + \varphi' + u$ with $(u, u) < \delta/3$ and also the expression after the lim sign in (61), with oppositely directed e_k becomes $-P\psi + u'$ with $(u', u') < \delta/3$. Then, on account of the unitary character of $T(a)$ and because of $T(-a) = T(a)^{-1}$

$$\Big(\psi, \sum_k \pm h^{-2}(T(2he_k) - 2T(he_k) + 1)\varphi\Big), = \Big(\sum_k \pm h^{-2}(T(-2he_k) - 2T(-he_k) + 1)\psi, \varphi\Big),$$

or

$$-P(\psi, \varphi) + (\psi, \varphi') + (\psi, u) = -P(\psi, \varphi) + (u', \varphi),$$

which is clearly impossible.

Thus if the lim in (61a) exists, it is $-P\varphi$ and this constitutes a characterization of the representation which is independent of similarity transformations. Since, according to the foregoing, it is always possible to find wave functions for a representation, to which (61a) can be applied, every reduced part of a representation with a given P must have this same P and no representation with one P can be contained in a representation with an other P. The same argument can be applied evidently to the positive and negative sub-classes of class 1 and 2.

B.

Every automorphism $L \to L^\circ$ of the group allows us to construct from one representation $D(L)$ another representation

$$D^\circ(L) = D(L^\circ). \tag{62}$$

This principle will allow us to restrict ourselves, for representations with finite, positive or negative P, to one value of P which can be taken respectively, to be $+1_+$ and -1. It will also allow in cases 1 and 2 to construct the representations of the negative sub-classes out of representations of the positive sub-classes.

The first automorphism is $a^\circ = \alpha a$, $\Lambda^\circ = \Lambda$. Evidently Equs. (12) are invariant under this transformation. If we set, however,

$$T^\circ(a)\varphi = T(\alpha a)\varphi; \qquad d^\circ(\Lambda)\varphi = d(\Lambda)\varphi,$$

then the occurring p

$$T^\circ(a)\varphi = T(\alpha a)\varphi = e^{i(p,\alpha a)}\varphi = e^{i(\alpha p, a)}\varphi,$$

will be the p occurring for the unprimed representation, multiplied by α. This allows, with a real positive α, to construct all representations with all possible numerical values of P, from all representation with one numerical value of P. If we take α negative, the representations of the negative sub-classes are obtained from the representations of the positive sub-class.

In case $P = 0_0$ evidently all representations go over into themselves by the transformation (62). In case $P = 0_+$ and $P = 0_-$ it will turn out that for positive α, (62) carries every representation into an equivalent one.

C.

On account of (53) and (56), (57), the equs. (52B) and (52C) are automatically satisfied and the $Q(p, \Lambda)_{\zeta\eta}$ must be determined by (52D). This gives

$$\sum_{\eta\vartheta} Q(p, \Lambda)_{\zeta\eta} Q(\Lambda^{-1}p, I)_{\eta\vartheta}\varphi(I^{-1}\Lambda^{-1}p, \vartheta) = \pm \sum_{\vartheta} Q(p, \Lambda I)_{\zeta\vartheta}\varphi(I^{-1}\Lambda^{-1}p, \vartheta). \tag{63}$$

Since this must hold for every φ, one would conclude

$$\sum_{\eta} Q(p, \Lambda)_{\zeta\eta} Q(\Lambda^{-1}p, I)_{\eta\vartheta} = \pm Q(p, \Lambda I)_{\zeta\vartheta}. \tag{63a}$$

Actually, this conclusion is not justified, since two wave functions must be considered to be equal even if they are different on a set of measure zero. Thus one cannot conclude, without further consideration, that the two sides of (63a) are equal at every point p. On the other hand,[24] the value of $Q(p, \Lambda)_{\zeta\eta}$ can be changed on a set of measure zero and one can make it continuous in the neighborhood of every point, if the representation is continuous. This allows then, to justify (63a). It follows from (63a) that $Q(p, 1)_{\zeta\eta} = \delta_{\zeta\eta}$.

Let us choose[15] now a basic p_0 arbitrarily. We can consider then the subgroup of all homogeneous Lorentz transformations which leave this p_0 unchanged. For all elements λ, ι of this "little group," we have

$$\sum_{\eta} Q(p_0, \lambda)_{\zeta\eta} Q(p_0, \iota)_{\eta\vartheta} = \pm Q(p_0, \lambda\iota)_{\zeta\vartheta}$$

$$q(\lambda)q(\iota) = \pm q(\lambda\iota), \tag{64}$$

where $g(\lambda)$ is the matrix $q(\lambda)_{\zeta\eta} = Q(p_0, \lambda)_{\zeta\eta}$. Because of the unitary character of $Q(\Lambda)$, the $Q(p_0, \Lambda)_{\zeta\eta}$ is unitary matrix and $q(\lambda)$ is unitary also.

If we consider, according to the last paragraph of Section 5, the group formed out of the translations and unimodular two-dimensional matrices, rather than Lorentz transformations, the $\pm$ sign in (64) can be replaced by a $+$ sign. In this case, λ and ι are unimodular two-dimensional matrices and the little group is formed by those matrices, the corresponding Lorentz transformations $\tilde{\lambda}$, $\tilde{\iota}$ to which leave p_0 unchanged $\tilde{\lambda}p_0 = \tilde{\iota}p_0 = p_0$.

Adopting this interpretation of (64), one can also see, conversely, that the representation $q(\lambda)$ of the little group, together with the class and P of the representation of the whole group, determines the latter representation, apart from a similarity transformation. In order to prove this, let us define for every p a two-dimensional unimodular matrix $\alpha(p)$ in such a way that the corresponding Lorentz transformation

$$\tilde{\alpha}(p)p_0 = p \tag{65}$$

brings p_0 into p. The $\alpha(p)$ can be quite arbitrary except of being an almost everywhere continuous function of p, especially continuous for $p = p_0$ and $\alpha(p_0) = 1$. Then, we can set

$$d(\alpha(p)^{-1})\varphi(p_0, \zeta) = \varphi(p, \zeta),$$

$$d(\alpha(p))\varphi(p, \zeta) = \varphi(p_0, \zeta). \tag{66}$$

This is equivalent to setting in (58)

$$Q(p, \alpha(p)) = 1 \tag{66a}$$

and can be achieved by a similarity transformation which replaces $\varphi(p, \zeta)$ by $\sum_{\eta} Q(p_0, \alpha(p)^{-1})_{\zeta\eta}\varphi(p, \eta)$. As the matrix $Q(p_0, \alpha(p)^{-1})$ is unitary, this is a unitary transformation. It does not affect, furthermore, (53) since it contains p only as a parameter.

Assuming this transformation to be carried out, (66) will be valid and will define, together with the $d(\lambda)$, all the remaining $Q(p, \Lambda)$ uniquely. In fact, calculating $d(\Lambda)\varphi(p, \zeta)$, we can decompose Λ into three factors

$$\Lambda = \alpha(p). \quad \alpha(p)^{-1}\Lambda\alpha(\tilde{\Lambda}^{-1}p). \quad \alpha(\tilde{\Lambda}^{-1}p)^{-1} \tag{67}$$

The second factor $\beta = \alpha(p)^{-1}\Lambda\alpha(\tilde{\Lambda}^{-1}p)$ belongs into the little group: $\tilde{\alpha}(p)^{-1}\tilde{\Lambda}\tilde{\alpha}(\tilde{\Lambda}^{-1}p)p_0 = \tilde{\alpha}(p)^{-1}\tilde{\Lambda}\cdot\tilde{\Lambda}^{-1}p = \tilde{\alpha}(p)^{-1}p = p_0$. We can write, therefore ($\tilde{\Lambda}^{-1}p = p'$)

$$\begin{aligned} d(\Lambda)\varphi(p,\zeta) &= d(\alpha(p))d(\beta)d(\alpha(p'))^{-1}\varphi(p,\zeta) \\ &= d(\beta)d(\alpha(p'))^{-1}\varphi(p_0,\zeta) \\ &= \sum_\eta q(\beta)_{\zeta\eta}d(\alpha(p')^{-1})\varphi(p_0,\eta) = \sum_\eta q(\beta)_{\zeta\eta}\varphi(p',\eta). \end{aligned} \tag{67a}$$

This shows that all representations of the whole inhomogeneous Lorentz group are equivalent which have the same P and the same representation of the little group. Further than this, the same holds even if the representations of the little group are not the same for the two representations but only equivalent to each other. Let us assume $q_1(\Lambda) = sq_2(\Lambda)s^{-1}$. Then by replacing $\varphi(p,\zeta)$ by $\sum_\eta s(\zeta,\eta)\varphi(p,\eta)$ we obtain a new form of the representation for which (53) still holds but $q_2(\beta)$ for the little group is replaced by $q_1(\beta)$. Then, by the transformation just described (Eq. (66)), we can bring $d(\Lambda)$ for both into the form (67a). The equivalence of two representations of the little group must be defined as the existence of a *unitary* transformation which transforms them into each other. (Only unitary transformations are used for the whole group, also).

On the other hand, if the representations of the whole group are equivalent, the representations of the little group are equivalent also: the representation of the whole group determines the representation of the little group up to a similarity transformation uniquely.

The representation of the little group was defined as the set of matrices $Q(p_0,\lambda)_{\zeta\eta}$ if the representation is so transformed that (53) and (66a) hold. Having two equivalent representations D and $SDS^{-1} = D^\circ$ for both of which (53) and (66a) holds, the unitary transformation S bringing the first into the second must leave all displacement operators invariant. Hence, it must have the form (57a), i.e., operate on the ζ only and depend on p only as on a parameter.

$$S\varphi(p,\zeta) = \sum_\eta S(p)_{\zeta\eta}\varphi(p,\eta). \tag{68}$$

Denoting the matrix Q for the two representations by Q and Q°, the condition $SD(\Lambda) = D^\circ(\Lambda)S$ gives that

$$\sum_\eta S(p)_{\zeta\eta}Q(p,\Lambda)_{\eta\vartheta} = \sum_\eta Q^\circ(p,\Lambda)_{\zeta\eta}S(\Lambda^{-1}p)_{\eta\vartheta} \tag{68a}$$

holds, for every Λ, for almost every p. Setting $\Lambda = \alpha(p_1)$ we can let p approach p_1 in such a way that (68a) remains valid. Since Q is a continuous function of p both $Q(p,\Lambda)$ and $Q^\circ(p,\Lambda)$ will approach their limiting value 1. It follows that there is no domain in which

$$S(p_1) = S(\alpha(p_1)^{-1}p_1) = S(p_0) \tag{69}$$

would not hold, i.e., that (69) holds for almost every p_1. Since all our equations must hold only for almost every p, the $S(p)_{\zeta\eta}$ can be assumed to be independent

of p and (68a) then to hold for every p also. It then follows that the representations of the little group in D and D° are transformed into each other by S_{t_τ}.

The definition of the little group involved an arbitrarily chosen momentum vector p_0. It is clear, however, that the little groups corresponding to two different momentum vectors p_0 and p are holomorphic. In fact they can be transformed into each other by $\alpha(p)$: If Λ is an element of the little group leaving p invariant then $\alpha(p)^{-1}\Lambda\alpha(p) = \beta$ is an element of the little group which leaves p_0 invariant. We can see furthermore from (67a) that if Λ is in the little group corresponding to p, i.e. $\Lambda p = p$ then the representation matrix $q(\beta)$ of the little group of p_0, corresponding to β, is identical with the representation matrix of the little group of p, corresponding to $\Lambda = \alpha(p)\beta\alpha(p)^{-1}$. Thus when characterizing a representation of the whole inhomogeneous Lorentz group by P and the representation of the little group, it is not necessary to say which p_0 is left invariant by the little group.

D.

Lastly we shall determine the constitution of the little group in the different cases.

1_+. In case 1_+ we can take for p_0 the vector with the components 0, 0, 0, 1. The little group which leaves this invariant obviously contains all rotations in the space of the first three coördinates. This holds for the little group of all representations of the first class.

0_0. In case 0_0, the little group is the whole homogeneous Lorentz group.

-1. In case $P = -1$ the p_0 can be assumed to have the components 1, 0, 0, 0. The little group then contains all transformations which leave the form $-x_2^2 - x_3^2 + x_4^2$ invariant, i.e., is the $2 + 1$ dimensional homogeneous Lorentz group. The same holds for all representations with $P < 0$.

0_+. The determination of the little group for $P = 0_+$ is somewhat more complicated. It can be done, however, rather simply, for the group of unimodular two dimensional matrices. The Lorentz transformation corresponding to the matrix $\begin{Vmatrix} a & b \\ c & d \end{Vmatrix}$ with $ad - bc = 1$ brings the vector with the components x_1, x_2, x_3, x_4, into the vector with the components x_1', x_2', x_3', x_4', where[18]

$$(70)\qquad \begin{Vmatrix} a & b \\ c & d \end{Vmatrix} \begin{Vmatrix} x_4 + x_3 & x_1 + ix_2 \\ x_1 - ix_2 & x_4 - x_3 \end{Vmatrix} \begin{Vmatrix} a^* & c^* \\ b^* & d^* \end{Vmatrix} = \begin{Vmatrix} x_4' + x_3' & x_1' + ix_2' \\ x_1' - ix_2' & x_4' - x_3' \end{Vmatrix}.$$

The condition that a null-vector p_0, say with the components 0, 0, 1, 1 be invariant is easily found to be $|a|^2 = 1, c = 0$. Hence the most general element of the little group can be written

$$(71)\qquad \begin{Vmatrix} e^{-i\beta/2} & (x + iy)e^{i\beta/2} \\ 0 & e^{i\beta/2} \end{Vmatrix},$$

with real x, y, β and $0 \leqq \beta < 4\pi$. The general element (71) can be written as $t(x, y)\ \delta(\beta)$ where

$$(71a) \qquad t(x, y) = \begin{Vmatrix} 1 & x + iy \\ 0 & 1 \end{Vmatrix}; \qquad \delta(\beta) = \begin{Vmatrix} e^{-i\beta/2} & 0 \\ 0 & e^{i\beta/2} \end{Vmatrix}.$$

The multiplication rules for these are

$$(71b) \qquad t(x, y)t(x', y') = t(x + x', y + y'),$$

$$(71c) \qquad \delta(\beta)t(x, y) = t(x \cos \beta + y \sin \beta, -x \sin \beta + y \cos \beta)\delta(\beta),$$

$$(71d) \qquad \delta(\beta)\delta(\beta') = \delta(\beta + \beta').$$

One could restrict the variability domain of β in $\delta(\beta)$ from 0 to 2π. As $\delta(2\pi)$ commutes with all elements of the little group, it will be a constant and from $\delta(2\pi)^2 = \delta(4\pi) = 1$ it can be $\delta(2\pi) = \pm 1$. Hence $\delta(\beta + 2\pi) = \pm\delta(\beta)$ and inserting a $\pm$ into equation (71d) one could restrict β to $0 \leqq \beta < 2\pi$.

These equations are analogous to the equations (52)–(52D) and show that the little group is, in this case, isomorphic with the inhomogeneous rotation group of two dimensions, i.e. the two dimensional Euclidean group.

It may be mentioned that the Lorentz transformations corresponding to $t(x, y)$ have elementary divisors, and constitute all transformations of class e) in 4B, for which $v_s = p_0$. The transformations $\delta(\beta)$ can be considered to be rotations in the ordinary three dimensional space, about the direction of the space part of the vector p_0. It is possible, then, to prove equations (71) also directly.

7. The Representations of the Little Groups

A. Representations of the three dimensional rotation group by unitary transformations.

The representations of the three dimensional rotation group in a space with a finite member of dimensions are well known. There is one irreducible representation with the dimensions 1, 2, 3, 4, $\cdots$ each, the representations with an odd number of dimensions are single valued, those with an even number of dimensions are two-valued. These representations will be denoted by $D^{(j)}(R)$ where the dimension is $2j + 1$. Thus for single valued representations j is an integer, for double valued representations a half integer. Every finite dimensional representation can be decomposed into these irreducible representations. Consequently those representations of the Lorentz group with positive P in which the representation of the little group—as defined by (64)—has a finite number of dimensions, can be decomposed into such representations in which the representation of the little group is one of the well known irreducible representations of the rotation group. This result will hold for all representations of the inhomogeneous Lorentz group with positive P, since we shall show that even the infinite dimensional representations of the rotation group can be decomposed into the same, finite, irreducible representations.

In the following, it is more appropriate to consider the subgroup of the two dimensional unimodular group which corresponds to rotations, than the rotation group itself, as we can restrict ourselves to single valued representations in this case (cf. equations (52)). From (70), one easily sees[18] that the condition for $\begin{Vmatrix} a & b \\ c & d \end{Vmatrix}$ to leave the vector with the components 0, 0, 0, 1 invariant is that it shall be unitary. It is, therefore, the two dimensional unimodular unitary group the representations of which we shall consider, instead of the representations of the rotation group.

Let us introduce a discrete coördinate system in the representation space and denote the coefficients of the unitary representation by $q(R)_{\kappa\lambda}$ where R is a two dimensional unitary transformation. The condition for the unitary character of the representation $q(R)$ gives

$$\sum_{\kappa} q(R)^*_{\kappa\lambda} q(R)_{\kappa\mu} = \delta_{\lambda\mu}; \qquad \sum_{\lambda} q(R)^*_{\kappa\lambda} q(R)_{\nu\lambda} = \delta_{\kappa\nu}, \tag{72}$$

$$\sum_{\kappa} |\, q(R)_{\kappa\lambda}\,|^2 = 1; \qquad \sum_{\lambda} |\, q(R)_{\kappa\lambda}\,|^2 = 1. \tag{72a}$$

This show also that $|\, q(R)_{\kappa\lambda}\,| \leqq 1$ and the $q(R)_{\kappa\lambda}$ are therefore, as functions of R, square integrable:

$$\int |\, q(R)_{\kappa\lambda}\,|^2\, dR$$

exists if $\int \cdots dR$ is the well known invariant integral in group space. Since this is finite for the rotation group (or the unimodular unitary group), it can be normalized to 1. We then have

$$\sum_{\kappa} \int |\, q(R)_{\kappa\lambda}\,|^2 dR = \sum_{\lambda} \int |\, q(R)_{\kappa\lambda}\,|^2 dR = 1. \tag{73}$$

The $(2j+1)^{\frac{1}{2}} D^{(j)}(R)_{kl}$ form,[26] a complete set of normalized orthogonal functions for R. We set

$$q(R)_{\kappa\lambda} = \sum_{jkl} C^{\kappa\lambda}_{jkl} D^{(j)}(R)_{kl}. \tag{74}$$

We shall calculate now the integral over group space of the product of $D^{(j)}(R)^*_{kl}$ and

$$q(RS)_{\kappa\mu} = \sum_{\lambda} q(R)_{\kappa\lambda} q(S)_{\lambda\mu}. \tag{75}$$

The sum on the right converges uniformly, as for (72a)

$$\sum_{\lambda=N}^{\infty} |\, q(R)_{\kappa\lambda} q(S)_{\lambda\mu}\,| \leqq \left(\sum_{\lambda=N}^{\infty} |\, q(R)_{\kappa\lambda}\,|^2 \sum_{\lambda=N}^{\infty} |\, q(S)_{\lambda\mu}\,|^2 \right)^{\frac{1}{2}} \leqq \left(\sum_{\lambda=N}^{\infty} |\, q(S)_{\lambda\mu}\,|^2 \right)^{\frac{1}{2}}$$

can be made arbitrarily small by choosing an N, independent of R, making the last expression small. Hence, (75) can be integrated term by term and gives

$$\int D^{(j)}(R)^*_{kl} q(RS)_{\kappa\mu} dR = \sum_{\lambda} \int D^{(j)}(R)^*_{kl} q(R)_{\kappa\lambda} q(S)_{\lambda\mu} dR. \tag{76}$$

[26] H. Weyl and F. Peter, Math. Annal. *97*, 737, 1927.

Substituting $\sum_m D^{(j)}(RS)_{km} D^{(j)}(S^{-1})_{ml}$ for $D^{(j)}(R)_{kl}$ one obtains

$$(77)\quad \sum_m D^{(j)}(S^{-1})^*_{ml} \int D^{(j)}(RS)^*_{km} q(RS)_{\kappa\mu} dR = \sum_\lambda q(S)_{\lambda\mu} \int D^{(j)}(R)^*_{kl} q(R)_{\kappa\lambda} dR.$$

In the invariant integral on the left of (77), R can be substituted for RS and we obtain, for (74) and the unitary character

$$(78)\quad \sum_m D^{(j)}(S)_{lm} C^{\kappa\mu}_{jkm} = \sum_\lambda q(S)_{\lambda\mu} C^{\kappa\lambda}_{jkl}.$$

Multiplying (78) by $D^{(h)}(S)^*_{in}$, the integration on the right side can be carried out term by term again, since the sum over λ converges uniformly

$$\sum_{\lambda=N}^{\infty} | C^{\kappa\lambda}_{jkl} q(S)_{\lambda\mu} | \leqq \left(\sum_{\lambda=N}^{\infty} | C^{\kappa\lambda}_{jkl} |^2 \sum_{\lambda=N}^{\infty} | q(S)_{\lambda\mu} |^2 \right)^{\frac{1}{2}} \leqq \left(\sum_{\lambda=N}^{\infty} | C^{\kappa\lambda}_{jkl} |^2 \right)^{\frac{1}{2}}.$$

This can be made arbitrarily small, as even $\sum_\lambda \sum_{jkl} (2j+1)^{-1} | C^{\kappa\lambda}_{jkl} |^2$ converges, for (74) and (72a). The integration of (78) yields thus

$$(79)\quad \sum_\lambda C^{\kappa\lambda}_{jkl} C^{\lambda\mu}_{hin} = \delta_{jh} \delta_{li} C^{\kappa\mu}_{jkn}.$$

From $q(R)q(E) = q(R)$ follows $q(E) = 1$ and then $q(R^{-1}) = q(R)^{-1} = q(R)^\dagger$. This, with the similar equation for $D^{(j)}(R)$ gives

$$(80)\quad \begin{aligned} \sum_{jkl} C^{\kappa\lambda}_{jkl} D^{(j)}(R^{-1})_{kl} &= q(R^{-1})_{\kappa\lambda} = q(R)^*_{\lambda\kappa} \\ &= \sum_{jkl} C^{\lambda\kappa *}_{jlk} D^{(j)}(R)^*_{lk} = \sum_{jkl} C^{\lambda\kappa *}_{jlk} D^{(j)}(R^{-1})_{kl}, \end{aligned}$$

or

$$(81)\quad C^{\kappa\lambda}_{jkl} = C^{\lambda\kappa *}_{jlk}.$$

On the other hand $q(E)_{\kappa\lambda} = \delta_{\kappa\lambda}$ yields

$$(82)\quad \sum_{jk} C^{\kappa\lambda}_{jkk} = \delta_{\kappa\lambda}.$$

These formulas suffice for the reduction of $q(R)$. Let us choose for every finite irreducible representation $D^{(j)}$ an index k, say $k = 0$. We define then, in the original space of the representation $q(R)$ vectors $v^{(\kappa jl)}$ with the components

$$C^{\kappa 1}_{jkl}, C^{\kappa 2}_{jkl}, C^{\kappa 3}_{jkl}, \ldots.$$

The vectors $v^{(\kappa jl)}$ for different j or l are orthogonal, the scalar product of those with the same j and l is independent of l. This follows from (79) and (81)

$$(83)\quad (v^{(\mu j' l')}, v^{(\kappa jl)}) = \sum_\lambda C^{\mu\lambda *}_{j'kl'} C^{\kappa\lambda}_{jkl} = \sum_\lambda C^{\kappa\lambda}_{jkl} C^{\lambda\mu}_{j'l'k} = \delta_{jj'} \delta_{ll'} C^{\kappa\mu}_{jkk}.$$

The $v^{(\kappa jl)}$ for all κ, j, l, form a complete set of vectors. In order to show this, it is sufficient to form, for every ν, a linear combination from them, the ν component of which is 1, all other components 0. This linear combination is

$$(84)\quad \sum_{\kappa jl} C^{\nu\kappa}_{jlk} v^{(\kappa jl)}.$$

In fact, the λ component of (84) is, on account of (79) and (82)

$$\sum_{\kappa jl} C^{\nu\kappa}_{jlk} C^{\kappa\lambda}_{jkl} = \sum_{jl} C^{\nu\lambda}_{jll} = \delta_{\nu\lambda}. \tag{85}$$

However, two v with the same j and l but different first indices κ are not orthogonal. We can choose for every j an l, say $l = 0$ and go through the vectors $v^{(1j0)}$, $v^{(2j0)}$, $\cdots$ and, following Schmidt's method, orthogonalize and normalize them. The vectors obtained in this way shall be denoted by

$$w^{(nj0)} = \sum_{\lambda} \alpha^{j}_{n\lambda} v^{(\lambda j0)}. \tag{86}$$

Then, since according to (83) the scalar products $(v^{(\kappa jl)}, v^{(\lambda jl)})$ do not depend on l, the vectors

$$w^{(njl)} = \sum_{\lambda} \alpha^{j}_{n\lambda} v^{(\lambda jl)} \tag{86a}$$

will be mutually orthogonal and normalized also and the vectors $w^{(njl)}$ for all n, j, l will form a complete set of orthonormal vectors. The same holds for the set of the conjugate complex vectors $w^{(njl)*}$. Using these vectors as coördinate axes for the original representation $q(R)$, we shall find that $q(R)$ is completely reduced. The ν component of the vector $q(R)v^{(\kappa jl)*}$ obtained by applying $q(R)$ on $v^{(\kappa jl)*}$ is

$$\sum_{\mu} q(R)_{\nu\mu} (v^{(\kappa jl)*})_{\mu} = \sum_{\mu} q(R)_{\nu\mu} C^{\mu\kappa}_{jlk}. \tag{87}$$

The right side is uniformly convergent. Hence, its product with $(2h + 1)$ $D^{(h)}(R)^{*}_{in}$ can be integrated term by term giving

$$\sum_{\mu} \int (2h+1) D^{(h)}(R)^{*}_{in} q(R)_{\nu\mu} C^{\mu\kappa}_{jlk}\, dR = \sum_{\mu} C^{\nu\mu}_{hin} C^{\mu\kappa}_{jlk} = \delta_{hj}\delta_{ln} C^{\nu\kappa}_{jik}. \tag{88}$$

Thus we have for almost all R

$$\sum_{\mu} q(R)_{\nu\mu} (v^{(\kappa jl)*})_{\mu} = \sum_{i} C^{\nu\kappa}_{jik} D^{(j)}(R)_{il} = \sum_{i} D^{(j)}(R)_{il} (v^{(\kappa ji)*})_{\nu}, \tag{88a}$$

or

$$q(R)v^{(\kappa jl)*} = \sum_{i} D^{(j)}(R)_{il} v^{(\kappa ji)*}. \tag{88b}$$

Since both sides are supposed to be strongly continuous functions of R, (88b) holds for every R. In (86a), for every n, the summation must be carried out only over a finite number of λ. We can write therefore immediately

$$q(R)w^{(njl)*} = \sum_{i} D^{(j)}(R)_{il} w^{(nji)*}. \tag{89}$$

This proves that the original representation decomposes in the coördinate system of the w into well known finite irreducible representations $D^{(j)}(R)$. Since the w form a complete orthonormal set of vectors, the transition corresponds to a unitary transformation.

This completes the proof of the complete reducibility of all (finite and infinite dimensional) representations of the rotation group or unimodular unitary group. It is clear also that the same consideration applies for all closed groups, i.e., whenever the invariant integral $\int dR$ converges.

The result for the inhomogeneous Lorentz group is: For every positive numerical value of P, the representations of the little group can be, in an irreducible representation, only the $D^{(0)}, D^{(\frac{1}{2})}, D^{(1)}, \cdots$, both for P_+ and for P_-. All these representations have been found already by Majorana and by Dirac and for positive P there are none in addition to these.

B. Representations of the two dimensional Euclidean group

This group, as pointed out in Section 6, has a great similarity with the inhomogeneous Lorentz group. It is possible, again[24], to introduce "momenta", i.e. variables ξ, η and v instead of the ζ in such a way that

$$t(x, y)\varphi(p_0, \xi, \eta, v) = e^{i(x\xi+y\eta)}\varphi(p_0, \xi, \eta, v). \tag{90}$$

Similarly, one can define again operators $R(\beta)$

$$R(\beta)\varphi(p_0, \xi, \eta, v) = \varphi(p_0, \xi', \eta', v), \tag{91}$$

where

$$\begin{aligned} \xi' &= \xi \cos\beta - \eta \sin\beta, \\ \eta' &= \xi \sin\beta + \eta \cos\beta. \end{aligned} \tag{91a}$$

Then $\delta(\beta)R(\beta)^{-1} = S(\beta)$ will commute, on account of (71c), with $t(x, y)$ and again contain ξ, η as parameter only. The equation corresponding to (57a) is

$$\delta(\beta)\varphi(p_0, \xi, \eta, v) = \sum_{\omega} S(\beta)_{v\omega}\varphi(p_0, \xi', \eta', \omega). \tag{92}$$

One can infer from (90) and (92) again that the variability domain of ξ, η can be restricted in such a way that all pairs ξ, η arise from one pair ξ_0, η_0 by a rotation, according (91a). We have, therefore two essentially different cases:

a.) $$\xi^2 + \eta^2 = \Xi \neq 0$$

b.) $$\xi^2 + \eta^2 = \Xi = 0, \quad \text{i.e.} \quad \xi = \eta = 0.$$

The positive definite metric in the ξ, η space excludes the other possibilities of section 6 which were made possible by the Lorentzian metric for the momenta, necessitated by (55).

Case b) can be settled very easily. The "little group" is, in this case, the group of rotations in a plane and we are interested in one and two valued irreducible representations. These are all one dimensional ($e^{is\beta}$)

$$S(\beta) = e^{is\beta} \tag{93}$$

where s is integer or half integer. These representations were also all found by Majorana and by Dirac. For $s = 0$ we have simply the equation $\Box\varphi = 0$,

for $s = \pm\frac{1}{2}$ Dirac's electron equation without mass, for $s = \pm 1$ Maxwell's electromagnetic equations, etc

In case a) the little group consists only of the unit matrix and the matrix $\begin{Vmatrix} -1 & 0 \\ 0 & -1 \end{Vmatrix}$ of the two dimensional unimodular group. This group has two irreducible representations, as (1) and (−1) can correspond to the above two dimensional matrix of the little group. This gives two new representations of the whole inhomogeneous Lorentz group, corresponding to every numerical value of Ξ. Both these sets belong to class 0_+ and two similar new sets belong to class 0_-.

The final result is thus as follows: The representations P_{+j} of the first subclass P_+ can be characterized by the two numbers P and j. From these P is positive, otherwise arbitrary, while j is an integer or a half integer, positive, or zero. The same holds for the subclass P_-. There are three kinds of representations of the subclass 0_+. Those of the first kind 0_{+s} can be characterized by a number s, which can be either an integer or a half integer, positive, negative or zero. Those of the second kind $0_+(\Xi)$ are single valued and can be characterized by an arbitrary positive number Ξ, those of the third kind $0'_+(\Xi)$ are double-valued and also can be characterized by a positive Ξ. The same holds for the subclass 0_-. The representations of the other classes (0_0 and P with $P < 0$) have not been determined.

8. Representations of the Extended Lorentz Group

A.

As most wave equations are invariant under a wider group than the one investigated in the previous sections, and as it is very probable that the laws of physics are all invariant under this wider group, it seems appropriate to investigate now how the results of the previous sections will be modified if we go over from the "restricted Lorentz group" defined in section 4A, to the extended Lorentz group. This extended Lorentz group contains in addition to the translations all the homogeneous transformations X satisfying (10)

$$XFX' = F \tag{10'}$$

while the homogeneous transformations of section 4A were restricted by two more conditions. From (10′) it follows that the determinant of X can be $+1$ or -1 only. If its -1, the determinant of $X_1 = XF$ is $+1$. If the four-four element of X_1 is negative, that of $X_2 = -X_1$ is positive. It is clear, therefore, that if X is a matrix of the extended Lorentz group, one of the matrices X, XF, $-X$, $-XF$ is in the restricted Lorentz group. For $F^2 = 1$, conversely, all homogeneous transformations of the extended Lorentz group can be obtained from the homogeneous transformations of the restricted group by multiplication with one of the matrices

$$1, F, -1, -F. \tag{94}$$

The group elements corresponding to these transformations will be denoted by E, F, I, IF. The restricted group contains those elements of the extended group which can be reached continuously from the unity. It follows that the transformation of an element L of the restricted group by F, I, or IF gives again an element of the restricted group. This is, therefore, an invariant subgroup of the extended Lorentz group. In order to find the representations of the extended Lorentz group, we shall use again Frobenius' method.[15]

We shall denote the operators corresponding in a representation to the homogeneous transformations (94) by $d(E) = 1, d(F), d(I), d(IF)$. For deriving the equations (52) it was necessary only to assume the existence of the transformations of the restricted group, it was not necessary to assume that these are the only transformations. These equations will hold, therefore, for elements of the restricted group, in representations of the extended group also. We normalize the indeterminate factors in $d(F)$ and $d(I)$ so that their squares become unity. Then we have $d(F)d(I) = \omega d(I)d(F)$ or $d(I) = \omega d(F)d(I)d(F)$. Squaring this, one obtains $\omega^2 = \pm 1$. We can set, therefore

$$\begin{aligned} d(IF) &= d(I)d(F) = \pm d(F)d(I) \\ d(F)^2 &= d(I)^2 = 1; \qquad d(IF)^2 = \pm 1. \end{aligned} \tag{95}$$

Finally, from

$$d(F)D(L_1)d(F) = \omega(L_1)D(FL_1F) \tag{96}$$

we obtain, multiplying this with the similar equation for L_2

$$\omega(L_1)\omega(L_2) = \omega(L_1L_2)$$

which, gives $\omega(L) = 1$ as the inhomogeneous Lorentz group (or the group used in (52B)-(52D)) has the only one dimensional representation by the unity (1). In this way, we obtain

$$d(F)D(L)d(F) = D(FLF), \tag{96a}$$

$$d(I)D(L)d(I) = D(ILI), \tag{96b}$$

$$d(IF)D(L)d(IF)^{-1} = D(IFLFI). \tag{96c}$$

B.

Given a representation of the extended Lorentz group, one can perform the transformations described in section 6A, by considering the elements of the restricted group only. We shall consider here only such representations of the extended group, for which, after having introduced the momenta, all representations of the restricted group are either in class 1 or 2, i.e. $P \geqq 0$ but not 0_0. Following then the procedure of section 6, one can find a set of wave functions for which the operators $D(L)$ of the restricted group have one of the forms, given in section 6 as irreducible representations. We shall proceed, next to find the operator $d(F)$. For the wave functions belonging to an irreducible $D(L)$ of the

restricted group, we can introduce a complete set of orthonormal functions $\psi_1(p, \zeta)$, $\psi_2(p, \zeta)$, $\cdots$. We then have

$$D(L)\psi_\kappa(p, \zeta) = \sum_\mu D(L)_{\mu\kappa}\psi_\mu(p, \zeta). \tag{97}$$

The infinite matrices $D(L)_{\mu\kappa}$ defined in (97) are unitary and form a representation which is equivalent to the representation by the operators $D(L)$. The $D(L)$, $d(F)$ are, of course, operators, but the $D(L)_{\mu\kappa}$ are components of a matrix, i.e. numbers. We can now form the wave functions $d(F)\psi_1$, $d(F)\psi_2$, $d(F)\psi_3$, $\cdots$ and apply $D(L)$ to these. For (96a) and (97) we have

$$\begin{aligned} D(L)d(F)\psi_\kappa &= d(F)D(FLF)\psi_\kappa \\ &= d(F)\sum_\mu D(FLF)_{\mu\kappa}\psi_\mu \\ &= \sum_\mu D(FLF)_{\mu\kappa}d(F)\psi_\mu . \end{aligned} \tag{97a}$$

The matrices $D^\circ(L)_{\mu\kappa} = D(FLF)_{\mu\kappa}$ give a representation of the restricted group (FLF is an element of the restricted group, we have a new representation by an automorphism, as discussed in section 6B). We shall find out whether $D^\circ(L)$ is equivalent $D(L)$ or not. The translation operation in D° is

$$T^\circ(a) = d(F)T(a)d(F) = T(Fa) \tag{98}$$

which, together with (53) shows that D° has the same P as $D(L)$ itself. In fact, writing

$$U_1\varphi(p, \zeta) = \varphi(Fp, \zeta) \tag{99}$$

one has $U_1^{-1} = U_1$ and one easily calculates $U_1T^\circ(a)U_1 = T(a)$. Similarly for $U_1d^\circ(\Lambda)U_1$ one has

$$\begin{aligned} U_1d^\circ(\Lambda)U_1\varphi(p, \zeta) &= U_1d(F\Lambda F)U_1\varphi(p, \zeta) \\ &= d(F\Lambda F)U_1\varphi(Fp, \zeta) = \sum_\eta Q(Fp, F\Lambda F)_{\zeta\eta}U_1\varphi(F\Lambda^{-1}p, \eta) \\ &= \sum_\eta Q(Fp, F\Lambda F)_{\zeta\eta}\varphi(\Lambda^{-1}p, \eta). \end{aligned} \tag{99a}$$

This means that the similarity transformation with U_1 brings $T^\circ(a)$ into $T(a)$ and $d^\circ(\Lambda)$ into $Q(Fp, F\Lambda F)P(\Lambda)$. Thus the representation of the "little group" in $U_1d^\circ(\Lambda)U_1$ is

$$q^\circ(\lambda) = Q(Fp_0, F\lambda F).$$

For this latter matrix, one obtains from (67a)

$$\begin{aligned} q^\circ(\lambda) = Q(Fp_0, F\lambda F) &= q(\alpha(Fp_0)^{-1}F\lambda F\alpha(Fp_0)) \\ &= q(\lambda^\circ) \end{aligned} \tag{100}$$

where λ° is obtained from λ by transforming it with $F\alpha(Fp_0)$.

The representations $D^\circ(L)$ and $D(L)$ are equivalent if the representation

$q(\lambda)$ is equivalent to the representation which coördinates $q(\lambda^\circ)$ to λ. The $\alpha(Fp_0)$ is a transformation of the restricted group which brings p_0 into $\alpha(Fp_0)p_0 = Fp_0$. (Cf. (65).) This transformation is, of course, not uniquely determined but if $\alpha(Fp_0)$ is one, the most general can be written as $\alpha(Fp_0)\iota$, where $\iota p_0 = p_0$ is in the little group. For $q(\iota^{-1}\alpha(Fp_0)^{-1}\,\Lambda\,\alpha(Fp_0)\iota) = q(\iota)^{-1}q(\alpha(Fp_0)^{-1}\,\Lambda\,\alpha(Fp_0))$ $q(\iota)$, the freedom in the choice of $\alpha(Fp_0)$ only amounts to a similarity transformation of $q^\circ(\lambda)$ and naturally does not change the equivalence or non equivalence of $q^\circ(\lambda)$ with $q(\lambda)$.

For the case P_+, we can choose p_0 in the direction of the fourth axis, with components 0, 0, 0, 1. Then $Fp_0 = p_0$ and $\alpha(Fp_0) = 1$. The little group is the group of rotations in ordinary space and $F\lambda F = \lambda$. Hence $q^\circ(\lambda) = q(\lambda)$ and $D^\circ(\Lambda)$ is equivalent to $D(\Lambda)$ in this case. The same holds for the representations of class P_-.

For 0_+ we can assume that p_0 has the components 0, 0, 1, 1. Then the components of Fp_0 are 0, 0, -1, 1. For $\alpha(Fp_0)$ we can take a rotation by π about the second axis and $F\alpha(Fp_0)$ will be a diagonal matrix with diagonal elements 1, -1, 1, 1, i.e., a reflection of the second axis. Thus if λ is the transformation in (70), $\lambda^0 = \alpha(Fp_0)^{-1}F\lambda F\alpha(Fp_0)$ is the transformation for which

$$\lambda^0\begin{pmatrix} x_4 + x_3 & x_1 - ix_2 \\ x_1 + ix_2 & x_4 - x_3 \end{pmatrix}\lambda^{0\dagger} = \begin{pmatrix} x_4' + x_3' & x_1' - ix_2' \\ x_1' + ix_2' & x_4' - x_3' \end{pmatrix}. \tag{101}$$

This is, however, clearly $\lambda^0 = \lambda^*$. Thus the operators of $q^0(\lambda)$ are obtained from the operators $q(\lambda)$ by (cf. (71a))

$$\begin{aligned} t^0(x, y) &= t(x, -y) \\ \delta^0(\beta) &= \delta(-\beta). \end{aligned} \tag{101a}$$

For the representations 0_{+s} with discrete s, the $q^0(\lambda)$ and $q(\lambda)$ are clearly inequivalent as $\delta^0(\beta) = (e^{-is\beta})$ and $\delta(\beta) = (e^{is\beta})$, except for $s = 0$, when they are equivalent. For the representations $0_+(\Xi)$, $0_+'(\Xi)$, the $q^0(\lambda)$ and $q(\lambda)$ are equivalent, both in the single valued and the double valued case, as the substitution $\eta \to -\eta$ transforms them into each other. The same holds for representations of the class 0_-. If $D^0(L)$ and $D(L)$ are equivalent

$$U^{-1}D^0(L)U = D(L), \tag{102}$$

the square of U commutes with all $D(L)$. As a consequence of this, U^2 must be a constant matrix. Otherwise, one could form, in well known manner,[27] an idempotent which is a function of U^2 and thus commutes with $D(L)$ also. Such an idempotent would lead to a reduction of the representation $D(L)$ of the restricted group. As a constant is free in U, we can set

$$U^2 = 1 \tag{102a}$$

[27] J. von Neumann, Ann. of Math. **32**. 191. 1931; ref. 2. p. 89.

restricted group, we can introduce a complete set of orthonormal functions $\psi_1(p, \zeta)$, $\psi_2(p, \zeta)$, $\cdots$. We then have

$$D(L)\psi_\kappa(p, \zeta) = \sum_\mu D(L)_{\mu\kappa}\psi_\mu(p, \zeta). \tag{97}$$

The infinite matrices $D(L)_{\mu\kappa}$ defined in (97) are unitary and form a representation which is equivalent to the representation by the operators $D(L)$. The $D(L)$, $d(F)$ are, of course, operators, but the $D(L)_{\mu\kappa}$ are components of a matrix, i.e. numbers. We can now form the wave functions $d(F)\psi_1$, $d(F)\psi_2$, $d(F)\psi_3$, $\cdots$ and apply $D(L)$ to these. For (96a) and (97) we have

$$\begin{aligned} D(L)d(F)\psi_\kappa &= d(F)D(FLF)\psi_\kappa \\ &= d(F)\sum_\mu D(FLF)_{\mu\kappa}\psi_\mu \\ &= \sum_\mu D(FLF)_{\mu\kappa}d(F)\psi_\mu . \end{aligned} \tag{97a}$$

The matrices $D^\circ(L)_{\mu\kappa} = D(FLF)_{\mu\kappa}$ give a representation of the restricted group (FLF is an element of the restricted group, we have a new representation by an automorphism, as discussed in section 6B). We shall find out whether $D^\circ(L)$ is equivalent $D(L)$ or not. The translation operation in D° is

$$T^\circ(a) = d(F)T(a)d(F) = T(Fa) \tag{98}$$

which, together with (53) shows that D° has the same P as $D(L)$ itself. In fact, writing

$$U_1\varphi(p, \zeta) = \varphi(Fp, \zeta) \tag{99}$$

one has $U_1^{-1} = U_1$ and one easily calculates $U_1T^\circ(a)U_1 = T(a)$. Similarly for $U_1d^\circ(\Lambda)U_1$ one has

$$\begin{aligned} U_1d^\circ(\Lambda)U_1\varphi(p, \zeta) &= U_1d(F\Lambda F)U_1\varphi(p, \zeta) \\ &= d(F\Lambda F)U_1\varphi(Fp, \zeta) = \sum_\eta Q(Fp, F\Lambda F)_{\zeta\eta}U_1\varphi(F\Lambda^{-1}p, \eta) \\ &= \sum_\eta Q(Fp, F\Lambda F)_{\zeta\eta}\varphi(\Lambda^{-1}p, \eta). \end{aligned} \tag{99a}$$

This means that the similarity transformation with U_1 brings $T^\circ(a)$ into $T(a)$ and $d^\circ(\Lambda)$ into $Q(Fp, F\Lambda F)P(\Lambda)$. Thus the representation of the "little group" in $U_1d^\circ(\Lambda)U_1$ is

$$q^\circ(\lambda) = Q(Fp_0, F\lambda F).$$

For this latter matrix, one obtains from (67a)

$$\begin{aligned} q^\circ(\lambda) = Q(Fp_0, F\lambda F) &= q(\alpha(Fp_0)^{-1}F\lambda F\alpha(Fp_0)) \\ &= q(\lambda^\circ) \end{aligned} \tag{100}$$

where λ° is obtained from λ by transforming it with $F\alpha(Fp_0)$.

The representations $D^\circ(L)$ and $D(L)$ are equivalent if the representation

$q(\lambda)$ is equivalent to the representation which coördinates $q(\lambda^\circ)$ to λ. The $\alpha(Fp_0)$ is a transformation of the restricted group which brings p_0 into $\alpha(Fp_0)p_0 = Fp_0$. (Cf. (65).) This transformation is, of course, not uniquely determined but if $\alpha(Fp_0)$ is one, the most general can be written as $\alpha(Fp_0)\iota$, where $\iota p_0 = p_0$ is in the little group. For $q(\iota^{-1}\alpha(Fp_0)^{-1}\Lambda\,\alpha(Fp_0)\iota) = q(\iota)^{-1}q(\alpha(Fp_0)^{-1}\Lambda\,\alpha(Fp_0))$ $q(\iota)$, the freedom in the choice of $\alpha(Fp_0)$ only amounts to a similarity transformation of $q^\circ(\lambda)$ and naturally does not change the equivalence or non equivalence of $q^\circ(\lambda)$ with $q(\lambda)$.

For the case P_+, we can choose p_0 in the direction of the fourth axis, with components 0, 0, 0, 1. Then $Fp_0 = p_0$ and $\alpha(Fp_0) = 1$. The little group is the group of rotations in ordinary space and $F\lambda F = \lambda$. Hence $q^\circ(\lambda) = q(\lambda)$ and $D^\circ(\Lambda)$ is equivalent to $D(\Lambda)$ in this case. The same holds for the representations of class P_-.

For 0_+ we can assume that p_0 has the components 0, 0, 1, 1. Then the components of Fp_0 are 0, 0, -1, 1. For $\alpha(Fp_0)$ we can take a rotation by π about the second axis and $F\alpha(Fp_0)$ will be a diagonal matrix with diagonal elements 1, -1, 1, 1, i.e., a reflection of the second axis. Thus if λ is the transformation in (70), $\lambda^0 = \alpha(Fp_0)^{-1}F\lambda F\alpha(Fp_0)$ is the transformation for which

$$\lambda^0\begin{pmatrix} x_4 + x_3 & x_1 - ix_2 \\ x_1 + ix_2 & x_4 - x_3 \end{pmatrix}\lambda^{0\dagger} = \begin{pmatrix} x_4' + x_3' & x_1' - ix_2' \\ x_1' + ix_2' & x_4' - x_3' \end{pmatrix}. \tag{101}$$

This is, however, clearly $\lambda^0 = \lambda^*$. Thus the operators of $q^0(\lambda)$ are obtained from the operators $q(\lambda)$ by (cf. (71a))

$$\begin{aligned} t^0(x, y) &= t(x, -y) \\ \delta^0(\beta) &= \delta(-\beta). \end{aligned} \tag{101a}$$

For the representations 0_{+s} with discrete s, the $q^0(\lambda)$ and $q(\lambda)$ are clearly inequivalent as $\delta^0(\beta) = (e^{-is\beta})$ and $\delta(\beta) = (e^{is\beta})$, except for $s = 0$, when they are equivalent. For the representations $0_+(\Xi)$, $0_+'(\Xi)$, the $q^0(\lambda)$ and $q(\lambda)$ are equivalent, both in the single valued and the double valued case, as the substitution $\eta \to -\eta$ transforms them into each other. The same holds for representations of the class 0_-. If $D^0(L)$ and $D(L)$ are equivalent

$$U^{-1}D^0(L)U = D(L), \tag{102}$$

the square of U commutes with all $D(L)$. As a consequence of this, U^2 must be a constant matrix. Otherwise, one could form, in well known manner,[27] an idempotent which is a function of U^2 and thus commutes with $D(L)$ also. Such an idempotent would lead to a reduction of the representation $D(L)$ of the restricted group. As a constant is free in U, we can set

$$U^2 = 1 \tag{102a}$$

[27] J. von Neumann, Ann. of Math. **32**. 191. 1931; ref. 2. p. 89.

C.

Returning now to equation (97a), if $D^0(L) = D(FLF)$ and $D(L)$ are equivalent ($P > 0$ or 0_+, 0_- with continuous Ξ or $s = 0$) there is a unitary matrix $U_{\mu\nu}$, corresponding to U, such that

$$\sum_\mu D(FLF)_{\kappa\mu} U_{\mu\nu} = \sum_\mu U_{\kappa\mu} D(L)_{\mu\nu}$$

$$\sum_\mu U_{\kappa\mu} U_{\mu\nu} = \delta_{\kappa\nu} . \tag{102b}$$

Let us now consider the functions

$$\varphi_\nu = \psi_\nu + \sum_\mu U_{\mu\nu} d(F) \psi_\mu . \tag{103}$$

Applying $D(L)$ to these

$$\begin{aligned} D(L)\varphi_\nu &= D(L)\psi_\nu + \sum_\mu U_{\mu\nu} D(L) d(F) \psi_\mu \\ &= D(L)\psi_\nu + \sum_\mu U_{\mu\nu} d(F) D(FLF) \psi_\mu \\ &= \sum_\mu D(L)_{\mu\nu} \psi_\mu + \sum_{\mu\kappa} U_{\mu\nu} d(F) D(FLF)_{\kappa\mu} \psi_\kappa \\ &= \sum_\mu D(L)_{\mu\nu} \Big(\psi_\mu + \sum_\kappa U_{\kappa\mu} d(F) \psi_\kappa \Big) = \sum_\mu D(L)_{\mu\nu} \varphi_\mu . \end{aligned} \tag{103a}$$

Similarly

$$\begin{aligned} d(F)\varphi_\nu &= d(F)\psi_\nu + \sum_\mu U_{\mu\nu} \psi_\mu \\ &= \sum_\mu U_{\mu\nu} \Big(\psi_\mu + \sum_\kappa U_{\kappa\mu} d(F) \psi_\kappa \Big) = \sum_\mu U_{\mu\nu} \varphi_\mu . \end{aligned} \tag{103b}$$

Thus the wave functions φ transform according to the representation in which $D(L)_{\mu\nu}$ corresponds to L and $U_{\mu\nu}$ to $d(F)$. The same holds for the wave functions

$$\varphi'_\nu = \psi_\nu - \sum_\mu U_{\mu\nu} d(F) \psi_\mu , \tag{104}$$

except that in this case $(-U_{\mu\nu})$ corresponds to $d(F)$. The ψ_ν and $d(F)\psi_\nu$ can be expressed by the φ and φ'. If the ψ and $d(F)\psi$ were linearly independent, the φ and φ' will be linearly independent also. If the $d(F)\psi$ were linear combinations of the ψ, either the φ or the φ' will vanish.

If we imagine a unitary representation of the group formed by the L and FL in the form in which it is completely reduced out as a representation of the group of restricted transformations L, the above procedure will lead to a reduction of that part of the representation of the group of the L and FL, for which $D(L)$ and $D(FLF)$ are equivalent.

If $D(L)_{\mu\nu}$ and $D^\circ(L)_{\mu\nu}$ are inequivalent, the ψ_κ and $d(F)\psi_\nu = \psi'_\nu$ are orthogonal. This is again a generalization of the similar rule for finite unitary

representations.[28] One can see this in the following way: Denoting $M_{\kappa\nu} = (\psi_\kappa, \psi'_\nu)$ one has

$$M_{\kappa\nu} = (\psi_\kappa, \psi'_\nu) = (D(L)\psi_\kappa, D(L)\psi'_\nu)$$
$$= \sum_{\mu\lambda} D(L)^*_{\mu\kappa} D^\circ(L)_{\lambda\nu} M_{\mu\lambda};$$
$$M = D(L)^\dagger M D^\circ(L).$$

Hence

$$D(L)M = MD^\circ(L); \qquad M^\dagger D(L) = D^\circ(L)M^\dagger. \tag{105}$$

From these, one easily infers that $MM^\dagger$ commutes with $D(L)$, and $M^\dagger M$ commutes with $D^\circ(L)$. Hence both are constant matrices, and if neither of them is zero, M and $M^\dagger$ are, apart from a constant, unitary. Thus $D(L)$ would be equivalent $D^\circ(L)$ which is contrary to supposition. Hence $MM^\dagger = 0$, $M = 0$ and the ψ are orthogonal to the $d(F)\psi = \psi'$. Together, they give a representation of the group formed by the restricted Lorentz group and F. If they do not form a complete set, the reduction can be continued as before.

One sees, thus, that introducing the operation F "doubles" the number of dimensions of the irreducible representations in which the little group was the two dimensional rotation group, while it does not increase the underlying linear manifold in the other cases. This is analogous to what happens, if one adjoins the reflection operation to the rotation groups themselves.[29]

D.

The operations $d(I)$ can be determined in the same manner as the $d(F)$ were found. A complete set of orthonormal functions corresponding to an irreducible representation of the group formed by the L and FL shall be denoted by ψ_1, $\psi_2, \cdots$. For this, we shall assume (97) again, although the $D(L)$ contained therein is now not necessarily irreducible for the restricted group alone but contains, in case of 0_{+s} or 0_{-s} and finite s, both s and $-s$. We shall set, furthermore

$$d(F)\psi_\kappa = \sum_\mu d(F)_{\mu\kappa}\psi_\mu. \tag{106}$$

We can form then the functions $d(I)\psi_1$, $d(I)\psi_2, \cdots$. The consideration, contained in (97a) shows that these transform according to $D(ILI)_{\mu\kappa}$ for the transformation L of the restricted group:

$$D(L)d(I)\psi_\kappa = \sum_\mu D(ILI)_{\mu\kappa} d(I)\psi_\mu. \tag{106a}$$

Choosing for L a pure translation, a consideration analogous to that performed in (98) shows that the set of momenta in the representation $L \to D(ILI)$ has the opposite sign to the set of momenta in the representation $D(L)$. If the latter

[28] Cf. e.g. E. Wigner, ref. 4, Chapter XII.
[29] I. Schur, Sitz. d. kön. Preuss. Akad. pages 189, 297, 1924.

belongs to a positive subclass, the former belongs to the corresponding negative subclass and conversely. Thus the adjunction of the transformation I always leads to a "doubling" of the number of states, the states of "negative energy" are attached to the system of possible states. One can describe all states ψ_1, ψ_2, $\cdots$, $d(I)\psi_1$, $d(I)\psi_2$, $\cdots$ by introducing momenta p_1, p_2, p_3, p_4 and restricting the variability domain of p by the condition $\{p, p\} = P$ alone without stipulating a definite sign for p_4.

As we saw before, the $d(I)\psi_1$, $d(I)\psi_2$, are orthogonal to the original set of wave functions ψ_1, ψ_2, $\cdots$. The result of the application of the operations $D(L)$ and $d(F)$ to the ψ_1, ψ_2, $\cdots$ (i.e., the representation of the group formed by the L, FL) was given in part C. The $D(L)d(I)\psi_\kappa$ are given in (106a). On account of the normalization of $d(I)$ we can set

$$d(I)d(I)\psi_\kappa = \psi_\kappa . \tag{106b}$$

For $d(F)d(I)\psi_\kappa$ we have two possibilities, according to the two possibilities in (95). We can either set

$$d(F)d(I)\psi_\kappa = d(I)d(F)\psi_\kappa = \sum_\mu d(F)_{\mu\kappa} d(I)\psi_\mu , \tag{107}$$

or

$$d(F)\cdot d(I)\psi_\kappa = -d(I)d(F)\psi_\kappa = -\sum_\mu d(F)_{\mu\kappa} d(I)\psi_\mu . \tag{107a}$$

Strictly speaking, we thus obtain two different representations. The system of states satisfying (107) could be distinguished from the system of states for which (107a) is valid, however, only if we could really perform the transition to a new coördinate system by the transformation I. As this is, in reality, impossible, the representations distinguished by (107) and (107a) are not different in the same sense as the previously described representations are different.

I am much indebted to the Wisconsin Alumni Research Foundation for their aid enabling me to complete this research.

MADISON, WIS.

ANNALS OF MATHEMATICS
Vol. 40, No. 1, January, 1939

MINIMAL SURFACES OF HIGHER TOPOLOGICAL STRUCTURE

BY JESSE DOUGLAS

(Received June 3, 1938)

TABLE OF CONTENTS

I. GEOMETRIC BACKGROUND AND STATEMENT OF RESULTS

1. Introduction 205
2. Representation of surfaces of general topological form 210
3. Orientation of the contours 218
4. Surfaces of assigned topological type bounded by given contours 224
5. Sufficient conditions for the solvability of the Plateau problem 228

II. ANALYTIC TREATMENT AND PROOFS

6. Symmetry and conjugacy 234
7. The Green's function of a symmetric Riemann surface 237
8. Solution of the Dirichlet problem for a general Riemann surface 243
9. The functional $A(g, R)$ 247
10. The identity 249
11. A special variation of the representation (g, R) 255
12. The first variation of $A(g, R)$ 259
13. Representations (g, R), proper and improper, as an L-set 262
14. Lower semi-continuity of $A(g, R)$ 267
15. Completion of the solution 273
16. Least area 276
17. Arbitrary Jordan contours 279
18. Conformal mapping of plane regions 287
19. Alternative method 294
20. Concluding remarks 297
21. Bibliography 297

CHAPTER I

GEOMETRIC BACKGROUND AND STATEMENT OF RESULTS

§1. INTRODUCTION

1

About two years ago, the author published a solution of the Plateau problem for minimal surfaces of general topological form: k contours, each with an assigned sense of description, given in n-dimensional space; prescribed genus h

or characteristic[1] r, and assigned character of orientability (two-sided or one-sided).[1a] The publication was in the form of two papers, the first of which[2] stated the results and outlined the methods used, while the second [2] gave details and proofs.

Subsequently, an alternative method was outlined by R. Courant [11], who has so far published the details for the case of genus zero [12].[3]

Some years ago, the author disposed of the important particular cases of a doubly-connected minimal surface with two given boundaries [3], and of a Möbius strip [4]. The most fundamental case of a simply-connected minimal surface bounded by a given arbitrary Jordan curve had been previously solved by the author [5], and by an alternative method, with rather less generality, by T. Radó [14].

Although the title of the paper [3] specified "two contours," the formulas of that paper were developed for the general case.[4] The treatment of the general problem given in [1], [2], was based on these formulas and on the theory of abelian integrals and θ-functions on a Riemann surface of any finite genus.

The present paper gives, first, a self-contained treatment of the problem, reviewing and considerably amplifying the essential features of the papers [1, 2, 3]. Further, it provides certain simple supplementary considerations,[4a] needed only in the case of characteristic greater than one, which serve to complete the presentation of the paper [2]. This is the case where "alter-symmetric" circuits[5] are present on the basic Riemann surface R, and the Green's function of R acquires certain simple complementary terms. The effect of these is easily traced,[6] with the same final result of the existence of the minimal surface. Many of the notions and formulas concerning Riemann surfaces, which are developed incidentally in the course of the analytic treatment, are new.

In our main theorem, II,[7] the existence of the minimal surface M_r is established solely on the basis of the functional $A(g, R)$ introduced by the author—closely related to the Dirichlet functional—without any reference to area or the use

[1] See the paragraph following (2.1) for definition.

[1a] This general form of the Plateau problem was first explicitly formulated by the author in Bull. Amer. Math. Soc., v. 36 (1930), p. 50, where also is broadly indicated the method of solution here finally elaborated.

[2] [1] of the bibliography at the end of this paper, references to which will be made by numbers in square brackets.

[3] An independent presentation of the same method, as far as the one-contour case is concerned was given by L. Tonelli [13]. According to Tonelli, this method coincides substantially with that used by the author throughout his work on the Plateau problem. The chief point of difference consists in the use of the vector Dirichlet functional in its original form (5.12), without transformation into $A(g, R)$.

[4] [3], p. 324.

[4a] (Added in proof.) These have already been published in the preliminary note [31] to the present paper.

[5] See §6.1.

[6] See §12.

[7] See §5.

of the theory of conformal mapping. In theorem I, the least area property of M_r is then proved by employing the conformal mapping of polyhedral surfaces. In §19 an alternative method of proving the combined theorems I and II is given, which uses this conformal mapping from the start. These theorems suppose the given contours capable of bounding some surface of finite area. Finally, §17 contains a proof of a previously announced theorem,[8] expressing the solution of the Plateau problem for the case where the contours are perfectly general Jordan curves. Solely to fix the ideas, we may suppose that the given contours do not intersect one another, but our theory applies with practically no change to the case where mutual intersections are permitted.[9]

The conformal mapping of multiply-connected plane regions, which, as throughout the author's work, is included in the Plateau problem as the case $n = 2$, is discussed in §18. The classic mapping theorem of Schottky is reestablished by the author's method, and combined with a proof of a topological correspondence between the boundaries. This new form of treatment of the Schottky theorem was stated by the author in [1],[10] proved in [2]. §18 of the present paper is concerned with the details of proof of a certain essential inequality, (18.5).

2

The θ-functions on the Riemann surface R enter through the expression of the Green's function of R in terms of these functions. In a future paper, I intend to give a simpler and more generally applicable treatment, which will use the Green's function in an intrinsic way without employing for it any explicit formula.[10a]

3

For definition of minimal surface, we adopt the formulas given for $n = 3$ by Weierstrass:

$$x_i = \Re F_i(w) \qquad (i = 1, 2, \cdots, n) \tag{1.1}$$

$$\sum_{i=1}^{n} F_i'^2(w) = 0 \qquad (w = u + iv). \tag{1.2}$$

These express only the first variational condition in the problem of the surface of least area with given boundaries, so that a minimal surface in this sense may or may not have the least area for its boundaries. The minimal surfaces whose existence is established in the present paper will actually have minimum area for their boundaries and topological form.

---

[8] [2], p. 108, Theorem IV.

[9] Cf. §3.3, also [8].

[10] As Theorem III of that paper.

[10a] (Added in proof.) The essential features of this alternative treatment have already been published in [32]. A detailed presentation will soon appear in the American Journal of Mathematics. See also [33].

Thus, analytically expressed, our problem is to find a Riemann surface R of appropriate topological type, and upon it n uniform harmonic functions $\Re F_i(w)$, which obey (1.2), and represent a surface bounded by preassigned contours. If E, F, G denote the fundamental quantities of the harmonic surface (1.1), then

$$\sum_{i=1}^{n} F_i'^2(w) = (E - G) - 2iF. \tag{1.3}$$

Hence condition (1.2) is equivalent to

$$E = G, \qquad F = 0, \tag{1.4}$$

expressing conformal representation of the harmonic surface on the Riemann surface R.

It is to be observed that only $\Re F_i(w)$ is required to be uniform; in general, $F_i(w)$ itself will have $p = r + k - 1$ pure imaginary periods corresponding to various circuits on the Riemann surface. These interpret themselves as translation periods of the adjoint minimal surface $x_i = \Im F_i(w)$. One may consider the relation between the catenoid and its adjoint surface, the helicoid, where $p = 1$ and there is a single translation period.

4

If the required minimal surface M_r be considered as having two faces, corresponding to opposite senses of the normal, then it becomes a closed surface, for the two surfaces are united at the boundaries. Regarded in this way, M_r can be represented on a closed Riemann surface R, whose genus, equal to that of the double-faced M_r, is seen to be $p = r + k - 1$. When the number k of contours is arbitrary, then even in case $r = 0$, the genus p of R is capable of taking arbitrary values. For this reason, nearly all the essential features of the theory for surfaces of general topological form are already present in the case $r = 0$, i.e., where the topological type is that of a sphere with k perforations. The rather secondary nature of the modifications that result from a higher topological structure, as will appear in the sequel, constitutes perhaps one of the advantages of the present method.

5

Illustrations of minimal surfaces of higher topological type are given in the following diagrams. All are bounded by single contours. In fig. 1a, the minimal surface has the form of a torus from which a rather large simply-connected region has been removed.[11] Fig. 1b shows a one-sided minimal surface of the form of a Möbius strip. Fig. 1c represents a standard form of surface of higher topological type;[12] here the genus $h = 2$. The contour is supposed to be almost in one plane. By increasing the number of double-bridges such

[11] Cf. [2], p. 122. We first gave this example in our invited address to the American Mathematical Society, Oct. 1932, published as [6] (where, however, fig. 1a is not included).

[12] [20], p. 153.

as A_1B_1, A_2B_2, we can produce a minimal surface of arbitrary genus. Fig. 1e is a standard representation[12] of a one-sided surface; here $r = 3$. The half-twisted strip S produces one-sidedness and contributes unity to the topological characteristic. More contours can be introduced arbitrarily by means of additional simple bridges, or by perforations in the surface. All these examples can be realized by the soap film experiments of Plateau.

The same contours also bound simply-connected minimal surfaces. Thus in 1a, we have a minimal surface consisting approximately of two horizontal circular discs joined by a vertical catenoidal strip.[13] There is a second simply-connected minimal surface, where the circular discs are vertical and the catenoidal strip horizontal. Fig. 1d indicates a simply-connected minimal surface

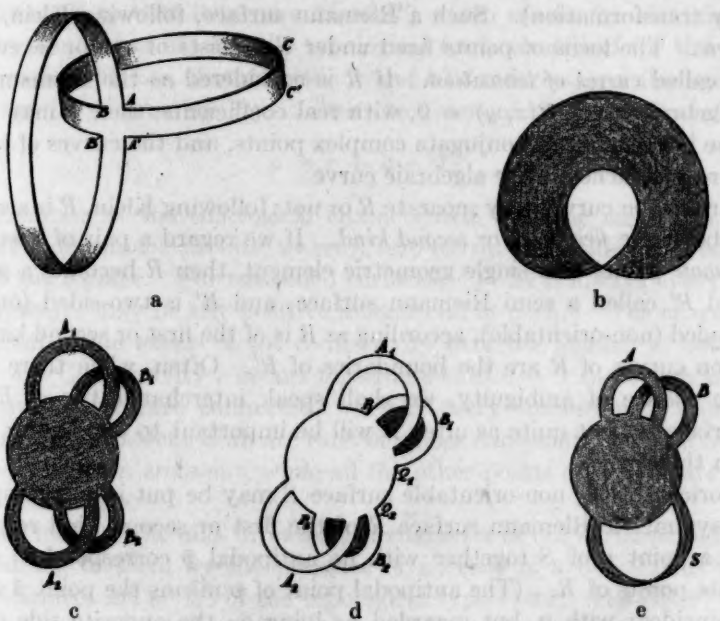


FIG. 1

bounded by the contour of fig. 1c. This consists, very approximately, of the circular disc + the heel-shaped regions $P_1B_1Q_1$, $P_2B_2Q_2$ + the strips A_1, A_2 + the shaded heel-shaped regions $\mathcal{H}_1$, $\mathcal{H}_2$, which are to be imagined slightly above the plane of the paper. We can also visualize a minimal surface of genus $h = 1$ with the same contour by performing the modifications indicated in fig. 1d on only one of the double-bridges. Supposing $\mathcal{H}_1 > \mathcal{H}_2$, we have, very approximately, in the notation (5.1):

$$a(\Gamma, 1) = a(\Gamma, 0) - 2\mathcal{H}_1, \tag{1.5}$$

$$a(\Gamma, 2) = a(\Gamma, 1) - 2\mathcal{H}_2, \tag{1.6}$$

[13] Cf. fig. 12b, with the narrow connecting part removed.

which illustrate the sufficient conditions of Theorem I (§5):

$$a(\Gamma, 2) < a(\Gamma, 1) < a(\Gamma, 0), \tag{1.7}$$

that assure the existence of the minimal surfaces described.

§2. Representation of Surfaces of General Topological Form

1

This section is based on the fundamental and highly suggestive tract of F. Klein, listed as [16] in the bibliography.[14]

Let R denote a Riemann surface having an inverse conformal transformation T into itself, where T associates the points of R in pairs, so that $T^2 = 1$ (involutory transformation). Such a Riemann surface, following Klein, is called *symmetric*. The locus of points fixed under T consists of one or several closed curves, called *curves of transition*. If R is considered as the Riemann surface of an algebraic curve $P(x, y) = 0$, with real coefficients, then T may be taken to be the interchange of conjugate complex points, and the curves of transition are the real branches of the algebraic curve.

The transition curves may separate R or not; following Klein, R is accordingly said to be of the *first kind* or *second kind*. If we regard a pair of T-equivalent or *conjugate* points as a single geometric element, then R becomes a geometric manifold R' called a semi Riemann surface, and R' is two-sided (orientable) or one-sided (non-orientable), according as R is of the first or second kind. The transition curves of R are the boundaries of R'. Often, when there seems to be little chance of ambiguity, we shall speak interchangeably of R and its semi-surface R', but quite as often it will be important to distinguish carefully between the two.[14a]

Any orientable or non-orientable surface S may be put into correspondence with a symmetric Riemann surface R of the first or second kind respectively, so that a point p of S together with its antipodal $\bar{p}$ correspond to a pair of conjugate points of R. (The antipodal point of p means the point $\bar{p}$ geometrically coincident with p, but regarded as lying on the opposite side of S, i.e., with reversed sense of the normal.) The distinction between the first and second kind of Riemann surface corresponds to the fact that on a two-sided surface we cannot pass from any point to its antipodal without crossing the boundary, whereas it is characteristic of a one-sided surface that such a passage is possible.

---

[14] Throughout this paper, the text-books [16–21] of the bibliography will be considered as current references.

[14a] Remark on notation: Thus, following our papers [1, 2], we use R in this paper to represent any complete, or closed, symmetric Riemann surface of finite genus, while R' denotes the corresponding semi-surface. This convention is followed as a rule with a few occasional exceptions, that need cause no real ambiguity. In subsequent papers [31–33], we have denoted the complete symmetric Riemann surface by an Old English or script $\mathfrak{R}$ and the related semi-surface by R.

In the form of the general symmetric Riemann surface R, we are provided with a complete system of standard domains of any finite genus or topological characteristic, with any finite number of boundaries, and either orientable or non-orientable. The completeness is from the standpoint of conformal mapping; i.e., the $3p - 3$ real conformal moduli (p = genus of R) are capable of varying arbitrarily in this system. From our point of view, conformally equivalent Riemann surfaces are identical, and one may replace the other if convenient for any purpose. The conformal equivalence must respect the symmetry of the surface; i.e., convert conjugate points into conjugate points. What is tantamount, the corresponding real algebraic curve $P(x, y) = 0$ is subject to real birational transformation, without change of anything essential to our discussion.

We choose R so that its semi-surface R', which results by identification of conjugate points, has exactly the topological form prescribed for the required minimal surface M, i.e., characteristic r and k boundaries, and also agrees with M in character of orientability. Then the genus of the *complete* surface R is

$$p = r + k - 1. \tag{2.1}$$

The definition of the topological characteristic r is: the maximum number of linearly independent circuits, where a zero circuit is considered as one which separates the surface. For two-sided surfaces, $r = 2h$ is always even; for one-sided surfaces, r may be any positive integer odd or even.[15] The characteristic is also equal to the excess of the connectivity over the number of boundaries: $c = r + k$. Connectivity c means that there exist $c - 1$ successive cross-cuts which leave the surface connected, whereas every succession of c cross-cuts disconnects. A cross-cut is an arc each of whose end-points lies on a boundary or previously drawn cross-cut, while all the other points of the arc are interior points of the surface.

We shall indicate the one- or two-sided character of our surfaces and quantities pertaining to them by the presence or absence of a bar. Thus we shall speak of surfaces M_r or $\bar{M}_{\bar{r}}$ of the topological type r or $\bar{r}$. Hereafter, we shall use r or $\bar{r}$ most frequently instead of the genus h, which applies only to orientable surfaces. Thus we may speak of the problem (Γ, r) or $(\Gamma, \bar{r})$, putting in evidence the set (Γ) of given contours and the topological type of the required minimal surface.

The following figure illustrates the case of a two-sided surface M, or symmetric Riemann surface R of the first kind, with $k = 4$, $r = 4$, therefore by (2.1), $p = 7$. Here the transition curves $C = C_1 + C_2 + C_3 + C_4$ separate R into two conjugate halves, and we may take R' to be the upper half. The inverse conformal transformation T may be pictured as the reflection in the plane containing C.

[15] Examples: Möbius strip, $r = 1$; with h handles, $r = 2h + 1$; Klein surface (see footnote 25), $r = 2$; with h handles, $r = 2h + 2$.

2

The one-sided case may be illustrated by the Riemann surface indicated in fig. 3. This is represented in the standard way as a double-sheeted surface over the z-plane. There are two branch-cuts, one C along the real axis, the other $q\bar{q}$ joining two conjugate complex points. We take the inverse conformal transformation T to be reflection in the real axis combined with passage to the other sheet. The points p $\bar{p}$ in the figure represent T-equivalent points, if we take p in the upper, $\bar{p}$ in the lower sheet. It is evident that T leaves invariant every point of the branch-cut C, and no other points. C is therefore the unique curve of transition; it is to be regarded as a closed curve, proceeding from a to b in one sheet and back to a in the other.

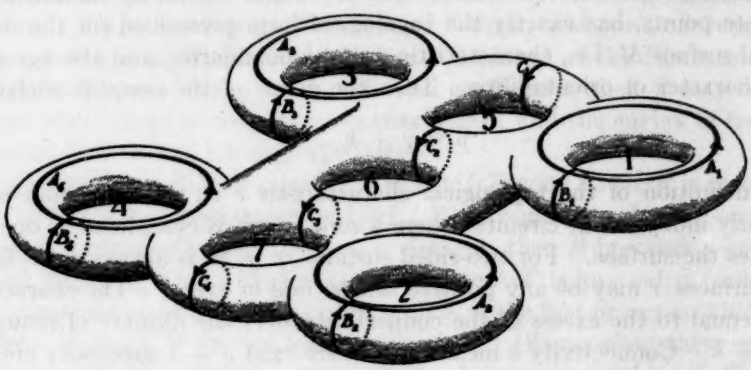


FIG. 2

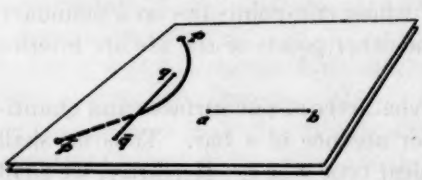


FIG. 3

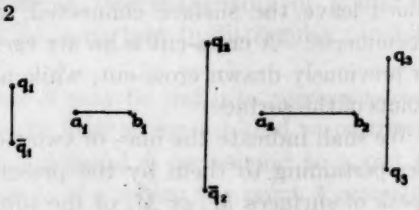


FIG. 4

Obviously, C does not separate R, for we can pass from p to $\bar{p}$ via the branch-cut $q\bar{q}$ without traversing C, as indicated in the figure. If we identify all conjugate point-pairs $p\bar{p}$, then R represents a Möbius strip bounded by C. Point-pairs such as $p\bar{p}$ correspond to a point on the Möbius strip together with its antipodal, and the one-sided nature of the strip is expressed exactly by the possibility of passing from p to $\bar{p}$ without crossing the boundary C.

By adding more branch-cuts like $q\bar{q}$ joining conjugate complex points, and branch-cuts like ab along the real axis, we can obtain models for one-sided surfaces with any characteristic r and any finite number k of boundaries. The following figure represents such a surface with two boundaries and characteristic three, where only the branch-cuts are indicated, along which the two sheets of the Riemann surface are joined.

More generally, we may illustrate both the orientable and non-orientable cases with Riemann surfaces of any number of sheets, which admit an inverse conformal transformation T into themselves, consisting of the conjugate transformation: z into $\bar{z}$, combined with a substitution S of period two ($S^2 = 1$) on the various sheets.

3

An alternative representation of one-sided surfaces $\bar{M}_r$ is based on the use of a two-sided covering surface, in the form of a semi Riemann surface $\mathfrak{R}'$ *of the first kind*[16] with an inverse conformal transformation U into itself, where U associates the points of $\mathfrak{R}'$ in pairs and leaves no point of $\mathfrak{R}'$ fixed. Each pair of U-equivalent points of $\mathfrak{R}'$ will correspond to the same point of $\bar{M}_r$. This one-two correspondence between $\bar{M}_r$ and $\mathfrak{R}'$ implies that the characteristic of $\mathfrak{R}'$ is $2r - 2$ and that $\mathfrak{R}'$ has $2k$ boundaries formed of k pairs of U-equivalent curves $C_1, C_1'; C_2, C_2'; \cdots; C_k, C_k'$. The genus of the *complete* Riemann surface $\mathfrak{R}$ of which $\mathfrak{R}'$ is one of the conjugate halves is

$$p = 2r + 2k - 3. \tag{2.2}$$

In any parametric representation of the k contours $\Gamma_1, \cdots, \Gamma_k$ on the transition curves C, each point of Γ_j corresponds to a pair of U-equivalent points on C_j, C_j' respectively.

For instance, a Möbius strip may be pictured in this way as a zone on the unit sphere $x^2 + y^2 + z^2 = 1$, bounded by the planes $z = \pm h$, where diametral points of the sphere are U-equivalent.[17]

Generally, we may picture $\mathfrak{R}$ as a Riemann surface of any number of sheets spread over the complex z-sphere and consisting of two conjugate halves $\mathfrak{R}', \mathfrak{R}''$ separated by the $2k$ transition curves $C_1, C_1'; \cdots; C_k, C_k'$. The inverse conformal transformation U of period two may be pictured as a diametral transformation $z' = -1/\bar{z}$ combined with a substitution of period two on the sheets which compose $\mathfrak{R}$. The transformation U converts each half $\mathfrak{R}', \mathfrak{R}''$ of $\mathfrak{R}$ into itself; besides, $\mathfrak{R}$ has an inverse conformal involution T which interchanges its two halves. T may be pictured as the reflection in the prime meridian plane, $z' = \bar{z}$, combined with a substitution of period two on the various sheets of $\mathfrak{R}$. The transformations U and T together generate a four-group $I, U, T, TU = UT$ of direct and inverse conformal transformations of $\mathfrak{R}$ into itself.

$\mathfrak{R}$ may be regarded as a covering surface in two-one correspondence with the previously used surface R (art. 2, this section), where each point of R corresponds to a pair of T-equivalent points of $\mathfrak{R}$, and U-equivalent points of $\mathfrak{R}$ correspond to conjugate points of R.

By means of the preceding representation, the Plateau problem $(\Gamma, \bar{r})$ is referred to $(\Gamma, -\Gamma, 2r - 2)$, where $-\Gamma$ denotes Γ with the reverse sense of description; i.e. we have $2k$ contours consisting of both orientations of each of

[16] That is, of the general type of the upper half of fig. 2.

[17] Cf. [4], p. 734.

the k given ones. We seek a semi Riemann surface $\mathfrak{R}'$ with a U-automorphism, $2k$ boundaries and characteristic $2r - 2$, and upon $\mathfrak{R}'$ a system of n harmonic functions $\Re F_i(w)$, *which take equal values at U-equivalent points*, obey $\sum_{i=1}^{n} F_i'^2(w) = 0$, and transform the boundaries of $\mathfrak{R}'$ in pairs into the opposite orientations of each contour Γ. Then $x_i = \Re F_i(w)$ or $\mathfrak{M}_{2r-2}$, will coincide with the desired one-sided surface $\bar{M}_r$ of type $\bar{r}$ bounded by (Γ). The two-sided surface $\mathfrak{M}_{2r-2}$ is, in fact, a covering surface in two-one correspondence with $\bar{M}_r$.

The simplest illustration is the case of the Möbius strip $(\Gamma_1, \bar{1})$, which we solved in the paper [4] by making it depend on the previously solved case [3] of a doubly-connected surface with two given contours: $(\Gamma_1, -\Gamma_1, 0)$. Here the two contours were the two opposite orientations of the single boundary assigned for the Möbius strip. The procedure there illustrated is completely typical of the general case.

By these remarks, we are enabled, as far as the analytic treatment is concerned, to confine ourselves to the two-sided case, with merely due regard to preserving the U-automorphism of $\mathfrak{R}$ when this Riemann surface is varied.

4. Reduction of R

If the symmetric Riemann surface R vary continuously, it may degenerate in various ways by the coalescence of branch-points, whereby certain branch-cuts disappear.[18] In this way the semi-surface R' may separate into a number of disconnected parts, or its characteristic r may be decreased. In case of separation, the total characteristic of the components is evidently $\leqq r$. We shall term either type of degeneration a *reduction* of R or R'.

Every reduction can be obtained by composition of primary reductions. A *primary reduction* of R consists of the following.

(1) In the two-sided case, the disappearance of a single pair of mutually symmetric branch-cuts.

(a) If the disappearing branch-cuts are the only ones which unite two given sheets, then R' separates into two disconnected parts: $R' = R_1' + R_2'$, with no decrease of the total characteristic: $r = r_1 + r_2$. The k transition curves C, or boundaries of R', are distributed between R_1' and R_2': $k = k_1 + k_2$. We may have $k_2 = 0$, in which case R_2' is a closed surface, and, as will be obvious, may simply be ignored for our purposes. If R_2' is closed, we shall not call the reduction primary unless $r_2 = 2$; if $r_2 = 0$, it may be considered that no reduction at all has taken place.

(b) If branch-cuts remain which join the same sheets as the disappearing ones, then the characteristic of R' is decreased by exactly *two*, to $r - 2$.

(2) In the one-sided case a primary reduction will mean:

(a) the disappearance of a single pair of mutually symmetric branch-cuts, *provided that this produces separation of* R': $R' = R_1' + R_2'$. Here again

[18] Alternatively, two branch-cuts may unite to form a single one, with the same effect.

$r = r_1 + r_2$. Then at least one of the components must be one-sided and both may be.

(b) the disappearance of a single self-symmetric branch-cut, i.e., one which joins a pair of conjugate branch-points. This reduces the characteristic by *one*, to $r - 1$. If $r - 1$ is even, the reduced surface may be two-sided or one-sided; if $r - 1$ is odd, it must be one-sided.

We shall denote primary reduction by an accent: r', $\bar{r}'$, and the general type of reduction by r'', $\bar{r}''$, interpreting the latter so as to include the former. In a composite reduction of a one-sided surface, there may be any distribution of one or two-sided components.

In terms of the corresponding algebraic equation $P(x, y) = 0$, the separation of R is equivalent to reducibility in the usual algebraic sense:

$$P(x, y) \equiv P_1(x, y)\, P_2(x, y) \cdots P_m(x, y). \tag{2.3}$$

The lowering of the genus of R corresponds to the acquisition of new double points produced by the coalescence of branch-points.

5. R' as a circular region

A useful alternative representation of a semi Riemann surface R' is as the fundamental region of a group L of linear fractional transformations of z or $\bar{z}$. The k boundaries of R' correspond to circles $C_1, C_2, \cdots, C_k$. If the genus $h = 0$, R' is simply the region bounded by these k circles, and L is generated by the inversions in them: $S_1, S_2, \cdots, S_k$. If R' is two-sided and $h > 0$, we must adjoin to L a set of h linear fractional transformations of z. This gives in the fundamental region h pairs of circles $K_1, K_1'; \cdots; K_h, K_{h'}$ whose points are coördinated by the corresponding linear fractional transformations. If R' is one-sided, of characteristic $r = 2h + 1$ or $2h + 2$, then we adjoin to $S_1, S_2, \cdots, S_k$ a set of h linear fractional transformations of z and, respectively, one or two linear fractional transformations T of $\bar{z}$ such that $T^2 = 1$. The latter contribute to the fundamental region their fixed circles, whose points are coördinated by T.

The complete symmetric Riemann surface R is obtained by adjoining to R' its inverse image in any one of the bounding circles $C_1, C_2, \cdots, C_k$, which form then the curves of transition.

For instance, a torus may be represented by a circular ring, on whose bounding circles radial points are coördinated. The group L is here generated by $z' = qz$, if the circles are about the origin with radii 1, q. Any number of boundaries on the torus may be represented by circles interior to the ring; we adjoin to L the inversions in these circles. A Möbius strip is represented by a circular ring, where diametral points of one of the bounding circles are coördinated. Equivalent to this is a zone on a sphere, where one of the bounding circles is a great circle on which diametral points are coördinated. Adjoining to this the diametral zone, we have a representation of the two-sided covering surface $\mathfrak{R}$

previously spoken of, on which U-equivalent points are diametral points and correspond to a single point of the Möbius strip.

This circular representation is quite simple and concrete, but we may observe the following advantages of the many-sheeted Riemann surface. The first consists in the fact that any separation of the Riemann surface is represented directly in a visible way, while in the circular region this separation corresponds to an approach to the same point of a number of circles. Thus in the latter case, only one of the components remains visible, the other disappears; in the former, both are on a par. This circumstance is important in defining certain basic functionals depending on R in the case when R reduces.

Again, in the Riemann surface representation, we have the choice of an extremely wide variety of conformally equivalent forms of a given surface, for the fundamental group is then the total real birational group, with an infinite number of parameters. In the circular representation, we have only the six-parameter group of transformations $z' = (az + b)/(cz + d)$. We can employ this greater freedom in the former case to avoid disadvantageous types of singularity, such as reduction to a point of a curve of transition, or real branch. For the group of real birational transformations is sufficiently large and effective to enable us to neutralize any such tendency by applying, roughly speaking, a kind of balancing magnification. In the circular representation any such attempt reduces some of the previously non-degenerating circles to points (sometimes at infinity).

For instance, in the two-contour case, we may use for R' the circular ring bounded by $C_1: |z| = 1$, $C_2: |z| = q < 1$. There are two forms of degeneration of R', namely $q = 0$, $q = 1$. In the former case, the circle C_2 disappears, in the latter the ring itself vanishes.

In contrast, we have the representation of R' as a semi-Riemann surface of a real algebraic curve of genus one with two real branches. A system of such Riemann surfaces, complete from the standpoint of conformal and symmetric equivalence, is represented, for instance, by the family of cubic curves:

$$(x^2 + y^2 - 1)(x - 2) = m,$$

$$0 < m < m_0 = \left(\frac{-2 + \sqrt{28}}{6}\right)\left(\frac{10 + \sqrt{28}}{6}\right)\left(\frac{8 - \sqrt{28}}{6}\right). \tag{2.4}$$

In this range of values of m, the curve has no double point; therefore, being cubic, it is of genus one. There are always two real branches, on which the two given contours may be represented parametrically.

The modulus m corresponds to the ratio of the radii q in the circular representation. At one extreme, $m = 0$, the curve is reducible, separating into a circle and a straight line, each of genus zero with a single real branch. This corresponds to $q = 0$. At the other extreme, $m = m_0$, the two real branches touch and the curve acquires a real double point. This corresponds to $q = 1$. We note that in both singular cases, the Riemann surface remains completely tangible in all respects. The figure shows the extreme curves $m = 0$, $m = m_0$

drawn full, and an intermediate one dotted. We may observe that in the representation as many-sheeted surface over the complex x-plane, one of the real branch-cuts or transition curves disappears, shrinking to the point $x = 2$, and the same happens to the corresponding component of the reducible surface. But if we pass to the representation over the y-plane—as we may do with preservation of conformality and symmetry—then there are two actual simply-connected components: one with branch-points at $y = \pm 1$, the other a simple y-plane, and the second transition curve becomes the entire real axis of this plane. This illustrates how easily undesirable types of degeneration may be avoided by conformal or birational transformation of the many-sheeted Riemann surface, where this is not possible for the circular representation.

The presence of infinite real branches can be avoided, if desired, by means of an inversion.

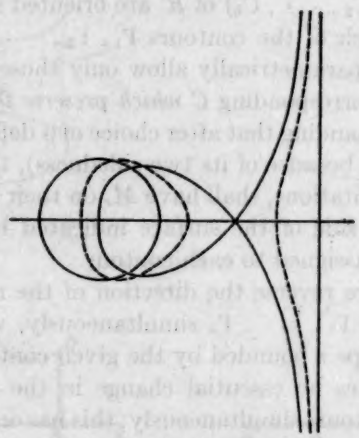


FIG. 5

Finally, if two of the contours $\Gamma_i\Gamma_j$ have a point in common, we can represent this in the circular form by two tangent circles C_iC_j. But, obviously, we cannot in this way represent more than two contours with the same point in common. On the other hand, in the representation by means of a Riemann surface or algebraic curve, we can arrange for real multiple points of any order of multiplicity.

For all these reasons, we shall employ in the sequel the representation by means of a Riemann surface. We may even restrict ourselves to the Clebsch-Lüroth normal form.[19] This is apart from a few occasions where the use of the circles seems more convenient, but in every case the situation can be recast in terms of Riemann surfaces.

Besides the many-sheeted form of R, we may consider the two-dimensional manifold S_2 which is the locus in real four-dimensional space x_1, x_2, y_1, y_2 of

[19] Cf. [18], pp. 440–444; also [19]. The essential feature of this normal form is that all branch-points are of first order, like that of $\sqrt{z}$.

the equation $P(x_1 + ix_2, y_1 + iy_2) = 0$. Incidentally S_2 is itself exactly a minimal surface in the four-space. Its orthogonal projection on the $x = x_1 + ix_2$ plane or the $y = y_1 + iy_2$ plane is a Riemann surface over that plane, on which S_2 is represented conformally by the projection.

§3. Orientation of the Given Contours

1. The 2^{k-1} different cases

The contours (Γ) must be given not only by themselves, but also each with a definite *sense* or *orientation*; in fact, a given set of k contours (Γ) gives rise to 2^{k-1} distinct forms of the Plateau problem, corresponding to that number of essentially different orientations of the set.

Consider first the two-sided case, of a minimal surface M_r. Suppose the boundaries $(C) = (C_1, C_2, \cdots, C_k)$ of R' are oriented so that R' is on the left. Then we may orient each of the contours $\Gamma_1, \Gamma_2, \cdots, \Gamma_k$ in a definite way, and in representing (Γ) parametrically allow only those topological representations of each Γ on the corresponding C *which preserve the sense*, as preassigned. This is equivalent to demanding that after choice of a definite sense of the normal of M_r (which is possible, because of its two-sidedness), the contours $\Gamma_1, \cdots, \Gamma_k$ in their preassigned orientations, shall have M_r on their left, from the viewpoint of one standing on the side of the surface indicated by the directed normal, and facing in the sense assigned to each contour.

It is evident that if we reverse the direction of the normal to M_r, and also that of *all* the contours $\Gamma_1, \cdots, \Gamma_k$ simultaneously, we still have a minimal surface of topological type r bounded by the given contours, and lying on their left. Therefore, it makes no essential change in the problem to reverse the sense of *all* the given contours simultaneously; this has only the effect of reversing the sense of the normal on the required minimal surface, and on all the surfaces whose areas are admitted to comparison.

However, if we change the sense on *some* but *not all* of the contours $\Gamma_1, \cdots, \Gamma_k$, then, whichever direction we choose for the normal to M_r, this surface no longer lies to the left of all the contours $\Gamma_1, \cdots, \Gamma_k$; i.e., it no longer furnishes a solution of the Plateau problem for the changed orientation of the given set of contours.

It is evident that we may keep fixed the sense on one of the contours Γ_1, but there are then exactly 2^{k-1} different forms of the Plateau problem connected with the given set (Γ), associated with that many possible combinations of sense on the other $k - 1$ contours $\Gamma_2, \cdots, \Gamma_k$. These 2^{k-1} problems are really quite different; for instance, a given set of contours (Γ) may bound a proper minimal surface of a prescribed topological type in one system of orientations but not in another.

The simplest illustration is $k = 2$, $r = 0$. Here we have two given contours Γ_1, Γ_2, and two essentially different Plateau problems, according to the sense which we assign to Γ_2 after that of Γ_1 has been fixed arbitrarily. Take, for

instance, two co-axial circles in planes whose distance apart is sufficiently small as compared with the radii. Then if the circles are sensed as in fig. 6a, the minimal surface which is bounded by them *on their left* exists, properly, being the catenoid with normal directed as in the figure. If, on the contrary, the circles are sensed as in fig. 6b, then no proper doubly-connected minimal surface exists which they bound on their left; the only possible solution is the pair of circular discs with normal directed upward.

If, for simplicity, we admit to comparison only surfaces of revolution, the difference between the two problems may be expressed as follows. In fig. 7a, the points P, Q are given on the *same* side of the X-axis; we require a curve joining these points, which, when revolved about the X-axis will generate a

FIG. 6

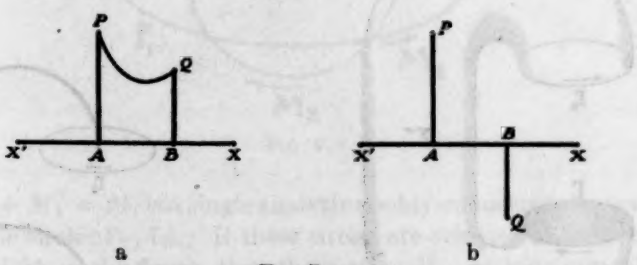


FIG. 7

minimum area. The solution is the catenary PQ. In fig. 7b, we have the same problem with the points P, Q' on *opposite* sides of the X-axis, the point P being in the same position as before, which Q' is the reflection of Q in the X-axis. Here the solution is the broken line $PABQ'$, as can easily be proved by elementary methods, with use of the observation that the length of any curve joining P and Q' is greater than $PA + Q'B$. It is evident that if we revolve about xx' any curve joining PQ or PQ' respectively, we obtain in either case a surface of revolution bounded by the same two circles, with respective centers A, B and radii PA, $QB = Q'B$.

The situation may also be expressed, quite concretely, as follows. Let us take a flexible and extensible tube, as in fig. 8a, whose ends are fixed to, say, two metal rings Γ_1, Γ_2 clamped in position in space. In this way a surface Σ

is formed bounded by Γ_1, Γ_2 and on their left, with the orientations of Γ_1, Γ_2 and the normal to Σ as in the figure. Now release the connection of the tube to Γ_2, keeping it fixed to Γ_1, and then stretch and bend the tube as in fig. 8b, bringing it down from above to be again attached to Γ_2. We obtain a surface Σ' bounded on the left by the same contours Γ_1, Γ_2 but in a different orientation: $+\Gamma_1$, $-\Gamma_2$.

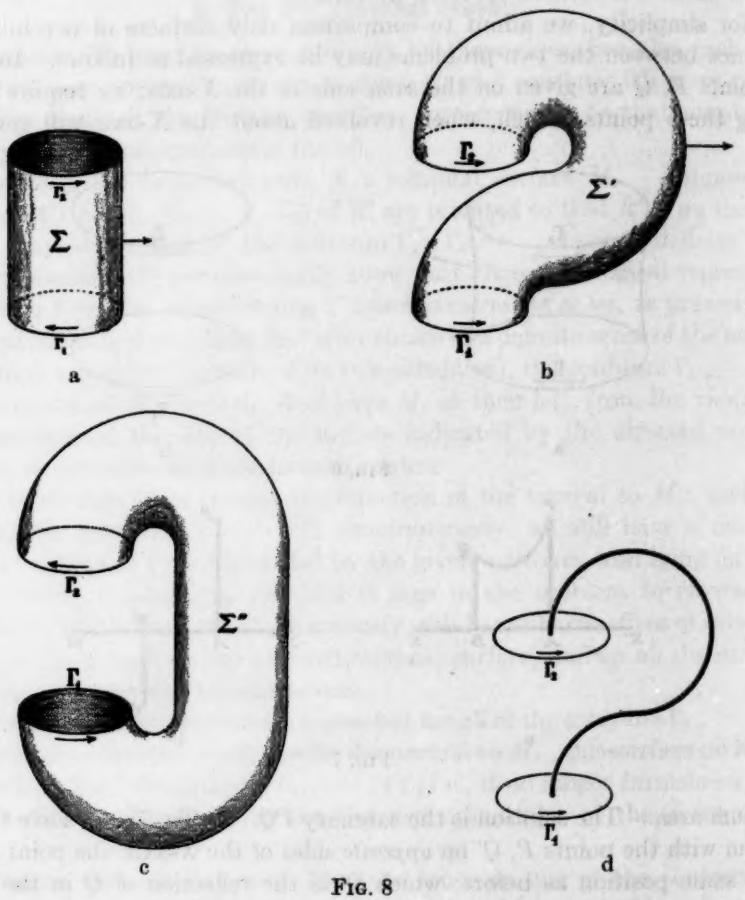


FIG. 8

Now we may put to ourselves two entirely distinct problems, namely:

(a) of all surfaces derivable from Σ by continuous deformation,[20] the boundaries Γ_1, Γ_2 remaining fixed, to find a minimal surface M of least area;

(b) the same for all surfaces derivable similarly from Σ'.

In the process of deformation, we may also allow the boundaries of Σ to vary, provided that they return to their original positions including sense. This

[20] Allowing any type of self-intersection in the process.

enables us to turn a surface "inside out," e.g., to pass from Σ to Σ'' (fig. 8c). The normal of Σ'' is directed towards the interior of the tube.

If Γ_1, Γ_2 are the two circles considered in fig. 6, then the solution of problem (a) is the catenoid, while that of problem (b) degenerates into the sum of the two circular discs.

Quite as often, however, we obtain two distinct proper solutions to the problems (a), (b). Consider, for example, two circles in perpendicular planes, with coincident centers and unequal radii. Form the single contour consisting of the line-segment AB, the upper semi-circle of Γ_2, the line segment CD, the forward semi-circle of Γ_1. This contour bounds a minimal surface of least area, M_1', lying entirely in the first quadrant, as we may call it. Let this surface be rotated through 180° about the diameter AD; then we get a minimal surface M_1'', lying in the third quadrant, bounded by AB, CD and the complementary semi-circles of Γ_1, Γ_2. According to the Schwarz symmetry principle for minimal surfaces,[21] M_1'' is the analytic prolongation of M_1' across AB and CD.

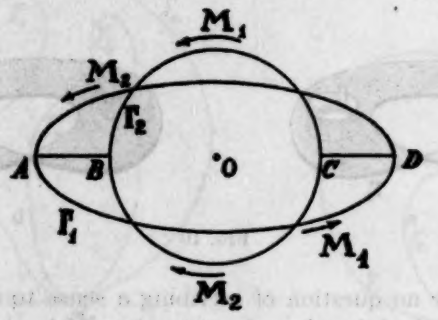


FIG. 9

Therefore $M_1' + M_1'' = M_1$ is a single analytic doubly-connected minimal surface, bounded by the circles Γ_1, Γ_2. If these circles are oriented as indicated by the arrows marked M_1 in the figure, then the surface M_1, with its normal properly directed, is on the left.

Let M_1 then be reflected in the plane of the vertical circle. We obtain a minimal surface M_2, also bounded by Γ_1, Γ_2, and while the sense of Γ_1 is preserved, the sense of Γ_2 must be reversed as indicated by the arrow marked M_2, in order to keep on the left the surface M_2 directed as obtained by the reflection from M_1.

If the vertical circle be moved to the right till it interlaces with the horizontal one, the preceding construction gives two (inversely congruent) doubly-connected minimal surfaces bounded by these interlacing contours. This illustrates

[21] [28], p. 175; [29], p. 397. Actually, a stronger form of this symmetry principle is necessary for the present purpose; see J. Douglas, *The analytic prolongation of a minimal surface over a rectilinear segment of its boundary*, Duke Mathematical Journal, Dec., 1938.

the general theorem given by the author in [3], p. 351, according to which any two interlacing contours bound a doubly-connected minimal surface. In the present example, contrary to what one might be inclined to guess, neither of the minimal surfaces is self-intersecting.

The reader will find it easy to illustrate the $2^2 = 4$ cases corresponding to the different orientations of three given contours.

2. One-sided surfaces

Consider next the case of one-sided surfaces. The figure illustrates the two ways in which we may interpret the problem of a surface of least area with characteristic one bounded by two given contours Γ_1, Γ_2 (Möbius surface with two boundaries). Here there is a proper solution only in the case of fig. 10a; in fig. 10b the solution degenerates to a Möbius strip Γ_1 and a circular disc Γ_2. However, if Γ_2, supposed to be a circle, is placed with its center on the Möbius strip determined by Γ_1, then, as I have proved, there exist proper one-sided solutions of characteristic one in *both* cases (a), (b).

FIG. 10

Here there can be no question of ascribing a sense to each contour so that the surface $\bar{M}$ is on the left, when the normal to $\bar{M}$ is properly directed, since there is no way of distinguishing *over the entire surface* $\bar{M}$ one direction of the normal from the other; indeed, this fact constitutes exactly the definition of a one-sided or non-orientable surface. The two senses of the normal are interchangeable when followed continuously over a properly arranged circuit.

We may say here that each contour occurs with *both* orientations. In the representation by means of the covering surface $\mathcal{R}'$ with k pairs of U-equivalent boundaries C_1, C_1'; $\cdots$; C_k, C_k' we have to decide which orientation of Γ_j shall go with C_j and which with C_j'. This gives again 2^{k-1} essentially distinct possibilities, since the first choice is indifferent.

3. Contours with common points

For definiteness, we shall suppose that the given contours (Γ) are k Jordan curves which do not interest one another. However, our theory and results are easily adapted to the case where the contours have points in common, provided no separate contours can be formed by combining arcs of the given ones. That is, we may have a situation like fig. 11a but not like fig. 11b.

All we need do in this case is to require the symmetric Riemann surface R to

have the system of its transition curves (C) topologically equivalent to the system (Γ) of given contours, that is, e.g. in fig. 11a, the corresponding algebraic curve $P(x, y) = 0$ must have exactly two real branches, one with a tripie point and a double point corresponding to Q_3 and Q_2 respectively, the other with a double point corresponding to Q_2'.

We have already shown in detail how this works out in the case of two contours with a single point in common, and where the minimal surface is required to be simply-connected.[22] The corresponding algebraic curve might be taken to be, e.g., a lemniscate, to whose real double point the common point Q of the two contours always corresponds. Actually, we used as semi Riemann surface a parallel strip of the z-plane, whose infinite point corresponded to the intersection point Q.[23]

An alternative method of regarding fig. 11a is to consider that we have two contours $\Gamma_1 + \Gamma_2 + \Gamma_3 + \Gamma_4$ and $\Gamma_5 + \Gamma_6$, which, however, are not Jordan curves,

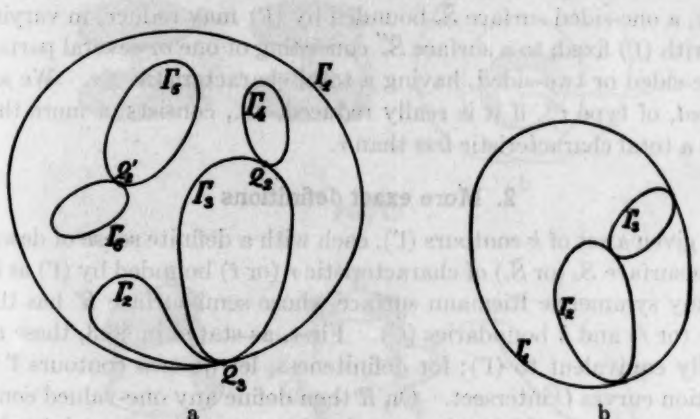


FIG. 11

but have multiple points. These can be taken account of by means of the sufficient condition (5.11) or (5.28), with the same final result as to the existence of the required minimal surface.

In short, we require the semi Riemann surface R' to be topologically equivalent to the required minimal surface M_r in all respects, including transition curves (C) as compared with given contours (Γ); we may say that R' is of type (Γ, r). If then, the k curves $\Gamma_1, \Gamma_2, \cdots, \Gamma_k$ which compose (Γ) have no multiple points either individually or mutually, then (C), as referred to the corresponding algebraic curve $P(x, y) = 0$, shall consist of k real branches without multiple points. On the other hand, in case the system of contours (Γ) presents multiple points, these must be represented as points of equal multiplicity on the system of real branches (C).

[22] [8].
[23] Cf. §10.2.

§4. Surfaces of Assigned Topological Type Bounded by Given Contours

1. Intuitive considerations

Let the surface S_r bounded by the contours (Γ) vary continuously in an arbitrary way, while the boundaries remain fixed. Then, in the limit, this surface may resolve itself into separate parts, or its topological characteristic may be reduced, e.g., by the closing up of an opening or by a handle of the surface shrinking to a curve Also a *closed* part of S_r may simply break off, giving a closed component without relation to the contours (Γ). These processes may occur several times successively or in any combination, so that we may obtain most generally as limit of a surface S_r bounded by (Γ) a surface S'' consisting of m parts having a total characteristic $\leqq r$. If either $m > 1$ or the characteristic is really $< r$, we say that S'' is of the *reduced* type r'' and write it S''_r.

Similarly, a one-sided surface $\bar{S}_r$ bounded by (Γ) may reduce, in varying continuously with (Γ) fixed, to a surface $\bar{S}''_r$ consisting of one or several parts, which may be one-sided or two-sided, having a total characteristic $\leqq r$. We say that $\bar{S}''_r$ is *reduced*, of type $\bar{r}''$, if it is really reduced, i.e., consists of more than one part or has a total characteristic *less* than r.

2. More exact definitions

Suppose given a set of k contours (Γ), each with a definite sense of description. We define a surface S_r (or $\bar{S}_r$) of characteristic r (or $\bar{r}$) bounded by (Γ) as follows. Let R be any symmetric Riemann surface whose semi-surface R' has the characteristic r (or $\bar{r}$) and k boundaries (C). First, as stated in §3.3, these must be topologically equivalent to (Γ); for definiteness, let no two contours Γ and no two transition curves C intersect. On R then define any one-valued continuous vector function $\mathfrak{x}(u, v)$ which takes equal values at conjugate points of R, and which transforms the transition curves (C) in a *monotonic continuous* way into (Γ). This means that the correspondence established by $\mathfrak{x}$ between (C) and (Γ) is either one-one continuous, or deviates from this, at most, by allowing any number[24] of *partial* arcs of one or more transition curves to correspond to single points on the related contours. The reference is to the depiction of the correspondence by a monotonic graph as in fig. 21. If the correspondence is one-one in any particular case, this will be especially proved.

Then the locus of the point $\mathfrak{x}(u, v)$ in the n-dimensional space will be called a *surface of characteristic* r *(or* $\bar{r}$*) bounded by* (Γ). This surface S_r (or $\bar{S}_r$) is two-sided or one-sided according as R is of the first or second kind in the Klein sense. The stipulation as to the coincidence of values of $\mathfrak{x}(u, v)$ at symmetric points makes $\mathfrak{x}$ a one-valued function on the semi-surface R', and we may regard $\mathfrak{x}$ as transforming R' topologically into S_r (or $\bar{S}_r$).

This definition gives S_r (or $\bar{S}_r$) in a particular *representation*. From this, its

[24] Finite or denumerably infinite.

most general representation may be derived by topological transformation of R into itself or a homeomorphic Riemann surface with preservation of conjugate points; in other words, the semi-surface R' undergoes topological transformation.

If R is reduced, of type r' or r'', then in case of separation of R, the continuity of $\mathfrak{x}(u, v)$ applies, of course, only to each component part. We have then defined a surface of the reduced type r' or r''—i.e., *primary reduced* or *general reduced*—bounded by (Γ). Similarly we have in the one-sided case reduced surfaces of type $\bar{r}'$, $\bar{r}''$ bounded by (Γ).

3. Reduction of surfaces

The simplest types of reduction are illustrated in the next figure. Fig. 12a illustrates a doubly-connected surface bounded by Γ_1, Γ_2 about to reduce to

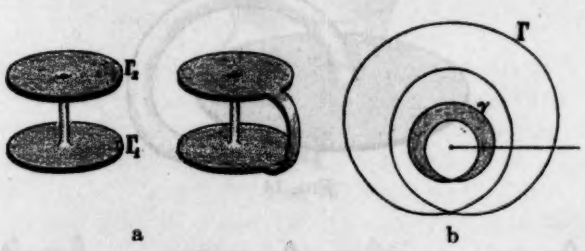


FIG. 12

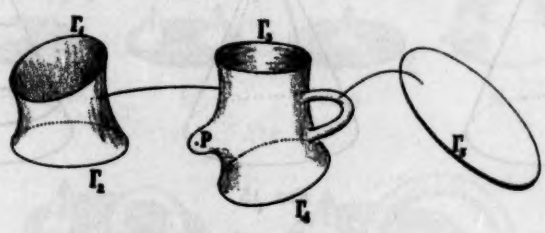


FIG. 13

the sum of two separate simply-connected ones. Fig. 12b shows a surface of torus type about to reduce into a simply-connected one. In fig. 12c, we start with the Riemann surface for $\sqrt{z}$ and delimit a branched portion by means of the contour Γ. Then we replace the branched part near the origin, bounded by γ, with a Möbius surface bounded by γ, as indicated. This gives a Möbius surface bounded by Γ. If now γ shrinks to the branch-point, this Möbius surface reduces back to the original simply-connected branched surface.

A more general type of reduction is illustrated in the next figure, which represents schematically the reduction of a surface S_6 bounded by *five* given contours into three separate parts with reduction of the total characteristic to *two*. One of the three handles, which gave the original characteristic *six*, has shrunk to a curve; the opening which produced another has shrunk to the point P; the third handle remains. The connections of the different parts have shrunk to curves,

which may simply be removed, as well as the withered handle. We may also imagine additional components in the form of closed surfaces attaching at isolated points to those pictured, which closed components may simply be broken off, without producing any new boundaries.

4. Reduction of one-sided surfaces

Starting with either a two-sided or a one-sided surface S_r or $\bar{S}_r$, we may convert it into a one-sided surface $\bar{S}_{r+2}$ with an increase of *two* in the characteristic, by attaching a handle joining opposite sides of S_r,[25] or simply any handle in the case of $\bar{S}_r$, already one-sided. Conversely, if this handle degenerates, $\bar{S}_{r+2}$ goes back to S_r or $\bar{S}_r$ with a reduction of *two* in the characteristic.

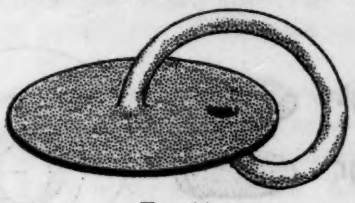


FIG. 14

FIG. 15

Often, however, it is desirable to reduce by exactly one unit the characteristic of a one-sided surface: $\bar{S}_r$ to $\bar{S}_{r-1}$ or S_{r-1}, and to do this by continuous deformation of $\bar{S}_r$.

We first consider, conversely, how to build S_{r-1} or $\bar{S}_{r-1}$ up to $\bar{S}_r$. This can be accomplished by the construction of a *cross-cap*,[26] as follows. We pinch up a small portion of S_{r-1} or $\bar{S}_{r-1}$ into the form of a cone. Then we slit this cone part way down along two elements, and join the opposite edges of the two slits cross-wise in the manner of a Riemann surface. The figure indicates the successive stages of the process. It is evident that by traversing the branch-cut, we

[25] Fig. 14 illustrates a Klein surface with a single boundary. A Klein surface is one obtained from a sphere by attaching a handle, one end of which is joined to the outside, the other to the inside of the spherical surface. A closed Klein surface is necessarily self-intersecting, in three dimensions. See [21].

[26] Kerékjártó [20], *Kreuzhaube*.

can pass from any point on the surface to its antipodal. Thus the surface has become one-sided, if it were not already so, and the characteristic has been increased by unity, corresponding to the addition of exactly one new circuit, leading from any point to its antipodal via the branch-cut.

Conversely, if the cross-cap disappears by continuous variation—the slit reducing to a point—the characteristic is reduced by unity, and the surface may be converted from one-sided to two-sided. This will happen if the one-sided nature is due solely to the cross-cap under consideration, and not also to topological phenomena elsewhere on the surface, i.e., other cross-caps or reverse handles.

FIG. 16

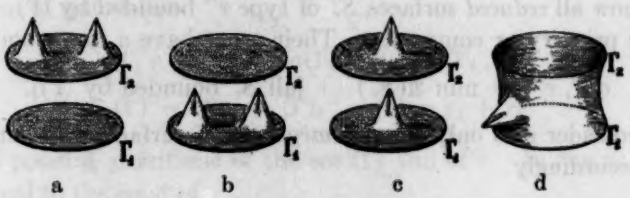


FIG. 17

FIG. 18

Fig. 16 illustrates, by means of cross-caps, a Klein surface with two given boundaries. The next figure shows all the forms of *primary* reduction. In order, these are

(a) simply-connected surface Γ_1 + Klein surface Γ_2,
(b) Klein surface Γ_1 + simply-connected surface Γ_2,
(c) Möbius surface Γ_1 + Möbius surface Γ_2,
(d) Möbius surface $\Gamma_1\Gamma_2$.

Fig. 18 illustrates two equivalent forms of a surface of type $\bar{3}$ with a single boundary. The surface being already rendered one-sided by the cross-cap, there is no distinction between a direct or reverse handle.

By disappearance of the cross-cap, we obtain from (a) a torus, from (b) a

Klein surface, with the given boundary. These represent the two possible types of primary reduction.

§5. Sufficient Conditions for the Solvability of the Plateau Problem

1. The functional $a(\Gamma, r)$

Consider all surfaces S_r which are *properly* of a given topological type r, and are bounded by a given set of oriented contours (Γ). Each of these has a definite area, which we understand to be defined in the *sense of Lebesgue*,[27] namely, as the least limit which can be approached by the area, in the elementary sense, of a variable polyhedral surface Π having the topological form of S_r and which tends to S_r. We then define:

$$a(\Gamma, r) = \min \mathfrak{A}(S_r), \qquad [\text{all } S_r \text{ bounded by } (\Gamma)]. \tag{5.1}$$

Throughout, we shall use "min" as synonymous with lower bound, without prejudice of the question as to whether the minimum is attained or not.

For any particular surface, $\mathfrak{A}(S_r)$ may be finite or infinite, and, as we shall see, there are even cases where $\mathfrak{A}(S_r)$ is identically infinite for given boundaries (Γ).[27a] Therefore $a(\Gamma, r)$ may be finite or $+\infty$.

Consider now all *reduced* surfaces S_r'' of type r'' bounded by (Γ), whether the reduction be primary or composite. Their areas have a lower bound

$$a(\Gamma, r'') = \min \mathfrak{A}(S_r'') \qquad [\text{all } S_r'' \text{ bounded by } (\Gamma)]. \tag{5.2}$$

We may consider also only the *primary reduced* surfaces S_r' bounded by (Γ), and define accordingly

$$a(\Gamma, r') = \min \mathfrak{A}(S_r') \qquad [\text{all } S_r' \text{ bounded by } (\Gamma)]. \tag{5.3}$$

The previous definitions apply also to one-sided surfaces: $\bar{r}$, $\bar{r}'$, $\bar{r}''$.

We have

$$a(\Gamma, r) \leqq a(\Gamma, r''). \tag{5.4}$$

For, given any reduced surface S_r'' bounded by (Γ), it is evident that by adjunction of appropriate connecting parts in the form of narrow tubes, of direct or reverse handles, and of cross-caps, *all having arbitrarily small area*, we can build up S_r'' to a *proper* surface S_r with the boundaries (Γ). From this consideration, (5.4) results immediately.

We may also express (5.4) in the form

$$a(\Gamma, r) = a(\Gamma, r, r''); \tag{5.5}$$

i.e., the lower bound of areas for the proper type r is also the lower bound for all of type r or r'', proper or reduced.

---

[27] [27], Lebesgue dealt explicitly only with simply-connected surfaces. Cf. formula (16.1).

[27a] See §17.5.

Evidently, in a similar manner to the passage from S_r'' to S_r, we can build up any composite reduced surface into one of primary type, with arbitrarily small modification of area; therefore

$$a(\Gamma, r') \leqq a(\Gamma, r''). \tag{5.7}$$

But since the set $[S_r']$ is part of the set $[S_r'']$, we have also the reverse inequality, and therefore

$$a(\Gamma, r') = a(\Gamma, r''). \tag{5.8}$$

This is the reason why, for all essential purposes, we may consider among reduced surfaces only the primary reduced type.

With (5.8), we have by (5.4),

$$a(\Gamma, r) \leqq a(\Gamma, r'). \tag{5.9}$$

Throughout, all definitions and theorems apply, with the appropriate change of wording and notation, also to one-sided surfaces: $\bar{r}$, $\bar{r}'$, $\bar{r}''$; i.e., we may interpret r as symbolizing: either r or $\bar{r}$.

By reference to §2.4, it appears that $a(\Gamma, r')$ is equal to the least of the following system of quantities:

$$a(\Gamma, r-2), \qquad a(\Gamma_1, r_1) + a(\Gamma_2, r_2), \tag{5.10}$$

where
$$(\Gamma) = (\Gamma)_1 + (\Gamma)_2, \qquad r = r_1 + r_2$$

represent all possible partitions of the set (Γ) and of r. In the one-sided case, $a(\Gamma, \bar{r}')$ is equal to the least of

$$a(\Gamma, \overline{r-1}), \qquad a(\Gamma, r-1), \tag{5.$\overline{10}$}$$
$$a(\Gamma_1, \bar{r}_1) + a(\Gamma_2, \bar{r}_2), \qquad a(\Gamma_1, \bar{r}_1) + a(\Gamma_2, r_2),$$

where only even values of the characteristic apply to any two-sided component.

2. Theorem I

Suppose that for given data (Γ, r) we really have in (5.9),

$$a(\Gamma, r) < a(\Gamma, r'). \tag{5.11}$$

Then we shall prove the following theorem.[27b]

THEOREM I. *If $a(\Gamma, r)$ is finite, and $a(\Gamma, r) < a(\Gamma, r')$, then a minimal surface M_r exists which is properly of topological type r and bounded by (Γ). The area of M_r is a minimum in comparison with all topologically equivalent surfaces S_r having the same boundaries (Γ), including sense. That is, $\mathfrak{A}(M_r) = a(\Gamma, r)$.*

[27b] First stated and proved in [1], [2].

3. The vector Dirichlet integral $D(\mathfrak{x})$

Although the interest and the concrete nature of the Plateau problem certainly depend on the interpretation as least area problem, we have found that from the standpoint of analytic treatment, it is in many ways more simple and effective to use the Dirichlet functional formed for a *vector* $\mathfrak{x}$:

$$D(\mathfrak{x}) = \int\int_{R'} \frac{1}{2}(E+G)\,dudv = \frac{1}{2}\int\int (\mathfrak{x}_u^2 + \mathfrak{x}_v^2)\,dudv \tag{5.12}$$
$$= \frac{1}{2}\int\int \sum_{i=1}^{n}\left[\left(\frac{\partial x_i}{\partial u}\right)^2 + \left(\frac{\partial x_i}{\partial v}\right)^2\right] dudv.$$

In fact, it may be said that the introduction of this functional to replace the area functional constitutes the main feature of the author's contribution in the solution of the Plateau problem.[27c]

To define $D(\mathfrak{x})$ in the one-sided case, the integration is performed over the entire Riemann surface R, and the effect of this neutralized by an extra factor $\frac{1}{2}$.

We suppose $\mathfrak{x}(u, v)$ to be continuously differentiable, at least piece-wise.

We shall make a distinction between the area $\mathfrak{A}(\mathfrak{x})$ in the sense of Lebesgue, and the area as defined by an integral:

$$I(\mathfrak{x}) = \int\int_{R'} \sqrt{EG - F^2}\,dudv. \tag{5.13}$$

The fact is that for surfaces $\mathfrak{x}$ of the stated regularity, the two are equal.[28] However, in the interest of a self-contained treatment, it is sufficient to know that certainly

$$\mathfrak{A}(\mathfrak{x}) \leqq I(\mathfrak{x}), \tag{5.14}$$

which is much more easily proved—see (16.11) and the associated discussion.

We observe that $\mathfrak{A}(\mathfrak{x})$ and $I(\mathfrak{x})$ depend only on the surface itself, being invariant under arbitrary parameter transformation, whereas $D(\mathfrak{x})$ is a functional of a surface in a given representation. The parameter transformation group under which $D(\mathfrak{x})$ is invariant is the conformal group, as is well-known.

Let us form the harmonic vector function $H(u, v)$ on R' with the same boundary values as $\mathfrak{x}(u, v)$ (Dirichlet problem). Then we have by the classic minimum property of harmonic functions in the Dirichlet integral,

$$D(H) \leqq D(\mathfrak{x}). \tag{5.15}$$

Evidently, then, the problem

$$D(\mathfrak{x}) = \text{min.}, \tag{5.16}$$

in the set

[all vectors $\mathfrak{x}(u, v)$ representing surfaces S_r bounded by (Γ)],

---

[27c] Cf. C. Carathéodory, *Bericht über die Verleihung der Fieldsmedaillen*, Comptes Rendus du Congrès International des Mathématiciens, Oslo, 1936, v. 1, pp. 308–314.

[28] See [15], pp. 11–12.

is entirely equivalent to the problem

$$D(H) = \text{min.}, \tag{5.17}$$

where $\mathfrak{x}$ is restricted to be a harmonic vector $H(u, v)$.

The harmonic function $H(u, v)$ depends only on the Riemann surface R and on the boundary values $g(z)$ of H on C, which form an arbitrary parametric representation of (Γ). Hence $D(H)$ is a functional of g and R, for which we use the notation

$$A(g, R) = D(H) = \frac{1}{2}\int\int_{R'} \sum_{i=1}^{n}\left[\left(\frac{\partial H_i}{\partial u}\right)^2 + \left(\frac{\partial H_i}{\partial v}\right)^2\right] du dv. \tag{5.18}$$

Therefore, equivalent to (5.16) and (5.17) is the problem

$$A(g, R) = \text{min.} \tag{5.19}$$

in the set $[(g, R)]$ consisting of all symmetric Riemann surfaces R whose semi-surfaces are of type (Γ, r), combined with all parametric representations g of (Γ) on the transition curves (C) of R.

Obviously, too,

$$\text{min } D(\mathfrak{x}) = \text{min } D(H) = \text{min } A(g, R) = d(\Gamma, r), \tag{5.20}$$

where the last symbol is the notation we now adopt for the common minimum value of these three aspects of the vector Dirichlet integral.

Reduced surfaces. Evidently, $D(\mathfrak{x})$ may be defined just as well for surfaces $\mathfrak{x}(u, v)$ of the reduced type r', r'', primary or general, as for the proper type r. We simply form, *by definition*, the sum of the integrals $D(\mathfrak{x})$ taken over the separate components of R'.

$H(u, v)$ with the boundary values $g(z)$, will be determined by solving a separate Dirichlet problem for each component part of R'. Accordingly, if

$$R = R_1 + R_2 + \cdots + R_m, \tag{5.21}$$

then, *by definition*,

$$A(g, R) = A(g_1, R_1) + A(g_2, R_2) + \cdots + A(g_m, R_m), \tag{5.22}$$

where g_j denotes the part of g that belongs to the transition curves of R_j. This is the reason that any closed components R_j' may be ignored, since a function harmonic and everywhere regular on a closed Riemann surface must reduce to a constant, and therefore contributes nothing to the Dirichlet integral.

We now define $d(\Gamma, r')$, $d(\Gamma, r'')$ as the common minimum value of $D(\mathfrak{x})$ or $D(H) = A(g, R)$, when the Riemann surface R varies over all of the reduced type r', r'' respectively, and g varies arbitrarily. Evidently, since type r'' includes r', we certainly have $d(\Gamma, r') \geqq d(\Gamma, r'')$, but the fact is, as will be proved, that

$$d(\Gamma, r') = d(\Gamma, r''), \tag{5.23}$$

which, as stated before in connection with area, justifies confining ourselves to the *primary* reduced type.

Dirichlet functional and area functional. We observe the identity:

$$\tfrac{1}{2}(E+G)-\sqrt{EG-F^2}=\tfrac{1}{2}(E^{\frac{1}{2}}-G^{\frac{1}{2}})^2 + \tfrac{1}{2}\{\sqrt{E^{\frac{1}{2}}G^{\frac{1}{2}}+F}-\sqrt{E^{\frac{1}{2}}G^{\frac{1}{2}}-F}\}^2. \tag{5.24}$$

This proves immediately:

THEOREM (a). $D(\mathfrak{x}) \geqq I(\mathfrak{x})$. *If* $D(\mathfrak{x}) = I(\mathfrak{x})$, *then* $E \equiv G$, $F \equiv 0$, *and conversely.*

In other words, $D(\mathfrak{x}) = I(\mathfrak{x})$ is the necessary and sufficient condition for a *conformal representation.*

By the first part of the theorem and (5.14), we certainly have

$$d(\Gamma, r) \geqq a(\Gamma, r). \tag{5.25}$$

Also, for reduced surfaces.

$$d(\Gamma, r') \geqq a(\Gamma, r'). \tag{5.25'}$$

The fact is, however, as will be proved, that

$$d(\Gamma, r) = a(\Gamma, r); \tag{5.26}$$

also

$$d(\Gamma, r') = a(\Gamma, r'). \tag{5.26'}$$

4. Theorem II

We shall also easily prove the analogue of (5.9):

$$d(\Gamma, r) \leqq d(\Gamma, r'). \tag{5.27}$$

Suppose that for given data (Γ, r), we really have the strong form of the inequality:

$$d(\Gamma, r) < d(\Gamma, r'). \tag{5.28}$$

This constitutes the sufficient condition in the statement of our main theorem.[28a]

THEOREM II. *If* $d(\Gamma, r)$ *is finite, and* $d(\Gamma, r) < d(\Gamma, r')$, *then a surface* M_r *bounded by* (Γ) *exists, in a representation* $M_r(u, v)$ *on a semi Riemann surface properly of topological type* r, *such that* $D(M_r(u, v)) = d(\Gamma, r)$. *Every such surface is minimal, and its representation is conformal.*

The spirit of our method consists in proving this theorem exclusively on the basis of the functional $A(g, R)$, without any reference to area, or the use of any part of the general theory of conformal mapping.

[28a] See footnote 27b.

5. Arbitrary Jordan contours

If the contours (Γ) are arbitrary Jordan curves, then in general

$$a(\Gamma, r) = +\infty, \tag{5.29}$$

i.e.

$$\mathfrak{A}(S_r) \equiv +\infty.$$

In order that $a(\Gamma, r)$ be finite, a necessary and sufficient condition is that *each* contour Γ_j be capable of bounding a *simply-connected* surface of finite area:

$$a(\Gamma_j) = a(\Gamma_j, 0) < +\infty, \qquad j = 1, 2, \cdots, k. \tag{5.30}$$

For this in turn a sufficient condition is, evidently, that each Γ_j be rectifiable; then, e.g., the cone obtained by joining all the points of Γ_j to any fixed point of space has finite area.

To prove the equivalence of $a(\Gamma, r) < +\infty$, and (5.30), we observe first that

$$a(\Gamma, r) \leqq a(\Gamma, r'') \leqq a(\Gamma_1) + a(\Gamma_2) + \cdots + a(\Gamma_k), \tag{5.31}$$

since the right member represents a particular type of reduction of S_r, namely, into k simply-connected parts. Therefore (5.30) implies that $a(\Gamma, r)$ is finite. Conversely, let $a(\Gamma, r)$ be finite; then a surface S_r exists bounded by (Γ) such that $\mathfrak{A}(S_r)$ is finite. We have then only to assume that for each contour Γ_j, it is possible to construct on S_r a rectifiable closed curve Γ_j' topologically equivalent to Γ_j by continuous variation *on* S_r. This is certainly the case if the defining function $\mathfrak{x}(u, v)$ of S_r is continuously differentiable piece-wise. Then the ring-shaped portion of S_r between Γ_j and Γ_j', combined with any conical surface having Γ_j' for base, evidently gives a simply-connected surface of finite area bounded by Γ_j.

Conversely stated, the necessary and sufficient condition for $a(\Gamma, r) = +\infty$ is that $a(\Gamma_j) = +\infty$ for at least one of the contours Γ_j.

For the case where $a(\Gamma, r)$ is *finite*, we define the functional

$$e(\Gamma, r) = a(\Gamma, r') - a(\Gamma, r) \geqq 0 \tag{5.32}$$

by (5.9).

For arbitrary Jordan curves Γ, whether capable of bounding surfaces of finite area or not, we define

$$\bar{e}(\Gamma, r) = \max \overline{\lim}\, e(\Gamma_m, r) \geqq 0, \tag{5.33}$$

the maximum being with respect to all possible ways of approaching to (Γ) with a sequence of contour systems $(\Gamma)_m$; $m = 1, 2, 3, \cdots$; such that each contour of each system is capable of bounding a finite area. This mode of approach is always possible, since the approaching contours may be taken to be, for instance, polygons.

Now, the following theorem holds.[28b]

[28b] Announced in [2], p. 108.

THEOREM III. *Suppose* $a(\Gamma, r) = +\infty$. *If* $\bar{e}(\Gamma, r) > 0$, *then a minimal surface* M_r *exists which is properly of topological type* r *and bounded by* (Γ). *The area of* M_r *is infinite, but every completely interior*[29] *portion* M' *has an area which is finite and a minimum for its own topological structure and boundaries.*

This theorem will be proved in §17.

In other words, our sufficient condition is:

(1) for the case $a(\Gamma, r)$ finite, that

$$a(\Gamma, r') - a(\Gamma, r)$$

be *strictly positive*;

(2) for the case $a(\Gamma, r) = +\infty$, that

$$\textit{the limit of } a(\Gamma, r') - a(\Gamma, r)$$

be strictly positive, for some way of tending to the given contour system (Γ) with a sequence of contour systems $(\Gamma)_m$ such that each approaching contour is finite-area-bounding.

Throughout this discussion we may replace the area functional by the Dirichlet functional and $a(\Gamma, r)$ by $d(\Gamma, r)$.

CHAPTER II

ANALYTIC TREATMENT AND PROOFS

§6. SYMMETRY AND CONJUGACY

1

Consider a symmetric Riemann surface R *of the first kind*, separated, that is, by the curves of transition. According to the remarks of §3, we can include in this way the one-sided as well as the two-sided case.

Fig. 2 will be used as a typical illustration.

If p denote the genus of R, there exist on R exactly $2p$ linearly independent circuits: p of the *first system* denoted by A_j, and p of the *second system* denoted by B_j. Of these, A_1, A_2, A_3, A_4 and B_1, B_2, B_3, B_4 are indicated in the figure; we can define $B_5 \sim C_1, B_6 \sim C_2, B_7 \sim C_3$ and A_5, A_6, A_7 may be easily visualized, surrounding the openings numbered correspondingly.

Suppose we give to A_5, A_6, A_7 a definite orientation, say the one which is counter-clockwise in the figure. Then we observe that these circuits are converted by T into themselves reversed:

$$(6.1') \qquad TA_5 = -A_5, \quad TA_6 = -A_6, \quad TA_7 = -A_7.$$

This leads us to arrange the orientations on A_1, A_3 and A_2, A_4 as in the figure, so that

$$(6.1'') \qquad TA_1 = -A_3, \quad TA_2 = -A_4, \quad TA_3 = -A_1, \quad TA_4 = -A_2.$$

---

[29] Meaning interior inclusive of its boundaries.

This orientation of the circuits of the first system determines, according to a standard convention of topology, a definite orientation of the circuits of the second system, as indicated in the figure. The convention is that the circuit equivalent to $\sum_1^p (A_j + B_j - A_j - B_j)$ shall have on its left the polygon of $4p$ sides to which R may be reduced by the canonical dissection. We then observe that under the transformation T, we have

$$\begin{aligned} TB_1 &= B_3, \quad TB_2 = B_4, \quad TB_3 = B_1, \quad TB_4 = B_2, \\ TB_5 &= B_5, \quad TB_6 = B_6, \quad TB_7 = B_7. \end{aligned} \tag{6.2}$$

The relations (6.1), (6.2) pair the circuits or indices j according to the substitution

$$T = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ 3 & 4 & 1 & 2 & 5 & 6 & 7 \end{pmatrix} \qquad (T^2 = 1). \tag{6.3}$$

We shall term *symmetric* to each other any two indices or circuits which correspond by this substitution, and shall use the notation

$$j' = \text{index symmetric to } j. \tag{6.4}$$

We may then summarize the relations (6.1′, 1″, 2) as follows:

$$TA_j = -A_{j'}, \quad TB_j = B_{j'}. \tag{6.5}$$

It is important to observe that there are two types of indices or circuits:

(1) *self-symmetric*: 5, 6, 7;

(2) *alter-symmetric*, which occur in pairs: (13), (24). The corresponding circuits B are self- or alter-symmetric according to their indices, while for the circuits A we have self- or alter-symmetry to within the sense, according to (6.1).

The alter-symmetric indices and circuits belong to the h handles of the semi-surface R' and their images by T on the conjugate semi-surface. *The essence of the case $h = 0$ (or $r = 0$) is that then all indices and circuits are self-symmetric, so that the substitution* (6.3) *induced by T reduces to the identity.*

The same applies to the case $r = 1$ of a Möbius surface with any number of boundaries. Here the covering semi-surface $\mathfrak{R}'$ is of genus zero.

But in the cases higher than this, $r > 1$, alter-symmetric circuits are actually present, and T induces an actual substitution of the indices, not the identity.

We shall use the notation:

$$\alpha, \beta, \lambda, \mu = 1, 2, \cdots, h \tag{6.6$_1$}$$

to denote indices belonging to alter-symmetric circuits on R';

$$\alpha', \beta', \lambda', \mu' = h + 1, h + 2, \cdots, 2h \tag{6.6$_2$}$$

to denote the *respective* symmetric indices belonging to circuits on the conjugate semi Riemann surface;

$$r, s = 1, 2, \cdots, k - 1 \tag{6.6$_3$}$$

to denote self-symmetric indices;

$$j, k, l, m = 1, 2, \cdots, p \tag{6.6$_4$}$$

to denote all indices indifferently. We shall also employ the standard summation convention, according to which the repetition of any index in a term denotes summation over the range of that index.

2

On the Riemann surface R there exist p *normal integrals of the first kind*: $v_1(w)$, $v_2(w)$, $\cdots$, $v_p(w)$. These can be continued indefinitely on R without coming to any singularity, but are not uniform; the period of v_j around A_k is $\delta_k^j (= 1$, if $j = k$; $= 0$, if $j \neq k)$, and its period as to B_k is denoted by τ_{jk}. The latter obey the well-known symmetry relation:

$$\tau_{jk} = \tau_{kj}. \tag{6.7}$$

Normal integrals v_j, $v_{j'}$ which correspond to symmetric indices will be called *symmetric* to each other. Between these we have the following important relations

$$v_j(w) = -\overline{v_{j'}(\overline{w})}, \tag{6.8}$$

where $\overline{w}$ denotes the symmetric point to w, and in general the bar denotes the conjugate complex quantity. This is for properly selected elements about w, $\overline{w}$ of the multiform functions v, and after the arbitrary additive constant involved in each v has been properly adjusted. The proof of (6.8) is by observing that the right side is everywhere regular, and has by (6.5) the same period about each circuit A_k as the left side. This identifies the two sides to within an additive constant, by the well-known theorem: an integral of first kind with periods of one system all equal to zero must reduce to a constant.

Allow w in (6.8) to describe B_k; then $\overline{w}$ describes $B_{k'}$, and we get

$$\tau_{jk} = -\bar{\tau}_{j'k'}. \tag{6.9}$$

This implies $\tau_{jj'}$ = pure imaginary, τ_{rs} = pure imaginary, with reference to the notation (6.6).

The quantities

$$t_{\alpha\beta} = \frac{1}{2\pi i}(\tau_{\alpha\beta} - \tau_{\alpha'\beta} - \tau_{\alpha\beta'} + \tau_{\alpha'\beta'}) \tag{6.10}$$

will play an important rôle. They are *real* by (6.9), in fact

$$t_{\alpha\beta} = \Re\frac{1}{\pi i}(\tau_{\alpha\beta} - \tau_{\alpha'\beta}) = \Re\frac{1}{\pi i}(\tau_{\alpha'\beta'} - \tau_{\alpha\beta'}). \tag{6.11}$$

Evidently, by (6.7), they are also symmetric in the indices α, β.

Their determinant $|\, t_{\alpha\beta} \,|$ can easily be proved non-vanishing. For the contrary would imply a system of linear relations

$$C_\alpha t_{\alpha\beta} = 0 \qquad (\beta = 1, 2, \cdots, h) \tag{6.12}$$

(all C_α real), where at least one $C_\alpha \neq 0$. Consider the everywhere regular harmonic function

$$H = C_\alpha \Re \frac{1}{\pi i} [v_\alpha(w) - v_{\alpha'}(w)]. \tag{6.13}$$

By (6.11, 12), its period about every circuit B_β, $B_{\beta'}$ vanishes. The same applies to every self-symmetric circuit B_r, since by (6.9),

$$\frac{1}{\pi i}(\tau_{\alpha r} - \tau_{\alpha' r}) = \frac{1}{\pi i}(\tau_{\alpha r} + \bar{\tau}_{\alpha r}) = \text{pure imaginary}. \tag{6.14}$$

Obviously too, the period of H about every circuit A_j is zero. Hence H is a *uniform* harmonic function without any singularity on the entire Riemann surface R; therefore H is a constant, and the same is consequently true of the complex analytic function of which it is the real part:

$$C_\alpha[v_\alpha(w) - v_{\alpha'}(w)] = C'. \tag{6.15}$$

Taking the period of each side about A_β, it follows that

$$C_\beta = 0 \text{ for } \beta = 1, 2, \cdots, h, \tag{6.16}$$

contrary to assumption.

Accordingly, $|\, t_{\alpha\beta} \,| \neq 0$, and we can therefore form the reciprocal matrix:

$$T_{\alpha\beta} = \text{cofactor of } t_{\alpha\beta} \div |\, t_{\alpha\beta} \,|. \tag{6.17}$$

Between $t_{\alpha\beta}$ and $T_{\alpha\beta}$, there are the standard relations

$$T_{\alpha\lambda} t_{\lambda\beta} = \delta_\beta^\alpha = \begin{cases} 1, & \text{if } \alpha = \beta, \\ 0, & \text{if } \alpha \neq \beta, \end{cases} \tag{6.18}$$

The reality and symmetry of $t_{\alpha\beta}$ imply the same properties for $T_{\alpha\beta}$.

§7. Green's Function of a Symmetric Riemann Surface

1. Definition

The Green's function $G(z, w)$ of a semi Riemann surface R' is defined uniquely by the following three properties:

(i) as a function of x, $y(z = x + iy)$, $G(z, w)$ is *uniform* and *harmonic* on R';

(ii) $G(z, w)$ is regular on R' except for a logarithmic singularity at $z = w$, so that for z in the neighborhood of w, we have

$G(z, w) = \log |\, z - w \,| +$ function regular at w;

(iii) on the transition curves C, which form the boundary of R', $G(z, w) = 0$.

It follows that by Schwarzian symmetry we can extend the function G over the entire Riemann surface R, and have

$$G(\bar{z}, w) = -G(z, w), \tag{iiia}$$

where z, $\bar{z}$ denote any two symmetric points of R. Conversely, property (iii) follows from (iiia), since on the transition curves, $\bar{z} = z$. Evidently, the function G, regarded thus as defined everywhere on R, is still uniform on R, and has,

besides $z = w$, also a logarithmic singularity at the symmetric point $\overline{w}$ with principal term $-\log | z - \overline{w} |$; otherwise G is regular on R.

We shall adopt this point of view as to $G(z, w)$, regarding it as defined everywhere on R, and determined uniquely by the conditions:

(I) $G(z, w)$ is *uniform* on R;

(II) $G(z, w)$ is *harmonic*;

(III) $G(z, w)$ is regular on R except for *two logarithmic singularities*, one at w with principal term $\log | z - w |$, the other at the symmetric point $\overline{w}$ with principal term $-\log | z - \overline{w} |$;

(IV) $G(\bar{z}, w) = -G(z, w)$.

2. The self-symmetric case, $r = 0, 1$

In the cases of self-symmetry, $r = 0, 1$, Green's function is expressed by the following simple explicit formula:[30]

$$G(z, w) = \Re\Omega(z; w, \overline{w}), \tag{7.1}$$

where Ω denotes the normal integral of the third kind with logarithmic singularities at w, $\overline{w}$.

The properties which characterize Ω to within an additive constant are:

(a) regular except for logarithmic singularities at w, $\overline{w}$ with respective principal terms $\log (z - w)$, $-\log (z - \overline{w})$;

(b) zero periods with respect to the circuits A of the first system.

An explicit formula for Ω in terms of θ-functions, which fixes also the additive constant, is

$$\Omega(z; w, \overline{w}) = \log \frac{\theta(v(z, w); \tfrac{1}{2}\beta, \tfrac{1}{2}\alpha, \tau)}{\theta(v(z, \overline{w}); \tfrac{1}{2}\beta, \tfrac{1}{2}\alpha; \tau)}. \tag{7.2}$$

Here θ denotes the multiple θ-series of Riemann, as generalized by Hermite,

$$\begin{aligned}\theta(u; \tau) = \sum \exp\, [&2\pi i u_j(n_j + \tfrac{1}{2}\alpha_j) \\ &+ i\pi\tau_{jk}(n_j + \tfrac{1}{2}\alpha_j)(n_k + \tfrac{1}{2}\alpha_k) + i\pi\beta_j(n_j + \tfrac{1}{2}\alpha_j)]\end{aligned} \tag{7.3}$$

The summation convention is employed for $j, k = 1, 2, \cdots, p$, and $\sum$ denotes a multiple summation as $n_1, n_2, \cdots, n_p$ range over all integers from $-\infty$ to $+\infty$. The characteristic $\frac{1}{2}\beta$, $\frac{1}{2}\alpha$ of the θ-series consists of $2p$ half-integers, which may be taken to be 0 or $\frac{1}{2}$; and we presume an *odd characteristic*, i.e.

$$\beta\alpha = \beta_1\alpha_1 + \beta_2\alpha_2 + \cdots + \beta_p\alpha_p = \text{odd integer}. \tag{7.4}$$

We shall also suppose

$$\alpha_{j'} = \alpha_j, \qquad \beta_{j'} = \beta_j; \tag{7.5}$$

in the self-symmetric case, $j' = j$, and this condition is satisfied of itself.

To form Ω, we substitute for the variables u_j the normal integrals of first kind, $v_j(z, w) = v_j(z) - v_j(w)$ in the numerator, and $v_j(z, \overline{w}) = v_j(z) - v_j(\overline{w})$

[30] [2], p. 109.

in the denominator. For the parameters τ_{jk}, we substitute the periods of these integrals, as defined in §6.

With the notation $\tau_{jk} = \tau'_{jk} + i\tau''_{jk}$, it is known[31] that the quadratic form $\tau''_{jk}(n_j + \frac{1}{2}\alpha_j)(n_k + \frac{1}{2}\alpha_k)$ is *positive definite*. This implies that the θ-series and all its derived series are absolutely and uniformly convergent for all bounded values of the variables; accordingly, it is permissible to differentiate this series term-wise to any order.

3. Properties of the Θ-function

The multiple θ-series has the following properties, important for the present application.

First, the partial differential equation

$$\frac{\partial^2\theta}{\partial u_j\,\partial u_k} = 4\pi i\,\frac{\partial\theta}{\partial\tau_{jk}}, \tag{7.6}$$

verifiable immediately by term-wise differentiation of the series (7.3).

Second, the period properties:

$$\theta(u_j + \delta^j_k) = e^{\pi i\alpha_k}\theta(u), \tag{7.7}$$

$$\theta(u_j + \tau_{jk}) = e^{-2\pi i(u_k + \frac{1}{2}\tau_{kk} + \frac{1}{2}\beta_k)}\theta(u). \tag{7.8}$$

We can verify (7.7) by direct substitution in the θ-series. To verify (7.8), we must accompany the indicated transformation of the variables u by a change of the summation indices: n_j into $n_j - \delta^j_k$. Throughout the calculation, of course, summation indices such as j, k may be freely interchanged.

Next, we have

$$\theta(-u) = e^{\pi i\beta\alpha}\theta(u). \tag{7.9}$$

In verifying this, we accompany the transformation of u into $-u$ by the change of summation indices: n_j into $-n_j - \alpha_j$. Under the condition of an odd characteristic, as expressed in (7.4), it follows that θ is an *odd function*:

$$\theta(-u) = -\theta(u). \tag{7.10}$$

Finally, we have the relation of conjugacy:

$$\theta(-\overline{u_{j'}}) = e^{\pi i\beta\alpha}\overline{\theta(u_j)}. \tag{7.11}$$

Unlike the preceding, this is dependent on the supposition (7.5), and the relations (6.9) of conjugacy which govern the periods τ_{jk}. In verifying (7.11), we make the transformation of summation indices: n_j into $n_{j'}$.

4. Properties of $\Omega(z; w, \overline{w})$

From (7.10) it follows that $\theta(0) = 0$. Hence the numerator in (7.2) vanishes when $z = w$, the denominator when $z = \overline{w}$. Each has exactly $p - 1$ other zeros

[31] [18], p. 83.

$a_1, a_2, \cdots, a_{p-1}$, which according to a known theorem on θ-functions, are the same for both.[32] Hence Ω has no other singularities except $w, \overline{w}$. Thus property (a) of Ω is implied.

Property (b) follows immediately from (7.7).

About any circuit B_k of the second system, we see by (7.8) that Ω has the period

$$2\pi i v_k(w, \overline{w}). \tag{7.12}$$

5. Properties of $G(z, w)$

The most important property of G is (I), its uniformity. Property (b) of Ω implies that the periods of G as to every circuit A vanish. The period of G as to any circuit B_k is, by (7.12),

$$\begin{aligned} P_k &= \Re 2\pi i v_k(w, \overline{w}) = \Re 2\pi i\,[v_k(w) - v_k(\overline{w})] \\ &= \Re\,[\text{pure imaginary quantity}] = 0, \end{aligned} \tag{7.13}$$

by (6.8), since $k' = k$ in virtue of the presumed self-symmetry of all indices. Hence G is *uniform* on R.

Property (a) of Ω implies properties (II) and (III) of G.

Finally, we easily verify, with the use of (7.11), that

$$\Omega(\bar{z}; w, \overline{w}) = -\overline{\Omega(z; w, \overline{w})}, \tag{7.14}$$

for properly arranged branches of the multiform function Ω. Taking real parts, we have

$$G(\bar{z}, w) = -G(z, w), \tag{7.15}$$

or property (IV) of G.

6. Alter-symmetric case, $r > 1$

In case alter-symmetric indices α are present, the formula $G = \Re\Omega$ no longer represents the Green's function of R. This is because $\Re\Omega$ is *no longer uniform* on R, due to the non-vanishing period

$$P_\alpha = \Re 2\pi i v_\alpha(w, \overline{w}) \tag{7.16}$$

with respect to each alter-symmetric circuit B_α. For we cannot apply here the reasoning associated with (7.13), which depends on $k' = k$, whereas $\alpha' \neq \alpha$.

The appropriate device for dealing with this situation is to modify $\Re\Omega$ by the addition of complementary terms so as to neutralize the periods P_α, and still preserve the other essential properties of Green's function. We have found the effective modification to be:[33]

$$G(z, w) = \Re\,\{\Omega(z; w, \overline{w}) - T_{\alpha\beta}\,[v_\alpha(z) - v_{\alpha'}(z)]\,[v_\beta(w, \overline{w}) - v_{\beta'}(w, \overline{w})]\}. \tag{7.17}$$

[32] [17], p. 262 (IV).

[33] As far as we know, an explicit formula for the Green's function of a general Riemann surface has not hitherto been given (except in the preliminary note [31] to the present paper).

We have

$$
\begin{aligned}
(7.18)\qquad v_\beta(w,\overline{w}) - v_{\beta'}(w,\overline{w}) &= v_\beta(w) - v_\beta(\overline{w}) - v_{\beta'}(w) + v_{\beta'}(\overline{w}) \\
&= v_\beta(w) - v_\beta(\overline{w}) + \overline{v_\beta(\overline{w})} - \overline{v_\beta(w)}, \qquad \text{by (6.8)}, \\
&= 2i\Im v_\beta(w,\overline{w}).
\end{aligned}
$$

Therefore, we may re-write (7.17):

$$
(7.19)\qquad G(z,w) = \Re\Omega(z;w,\overline{w}) - T_{\alpha\beta}\Re 2i\,[v_\alpha(z) - v_{\alpha'}(z)]\Im v_\beta(w,\overline{w}).
$$

Hence the period about B_λ of G is

$$
\begin{aligned}
P_\lambda &= \Re 2\pi i v_\lambda(w,\overline{w}) - T_{\alpha\beta}\Re 2i(\tau_{\alpha\lambda} - \tau_{\alpha'\lambda})\Im v_\beta(w,\overline{w}) \\
&= \Re 2\pi i v_\lambda(w,\overline{w}) + 2\pi T_{\alpha\beta} t_{\alpha\lambda}\Im v_\beta(w,\overline{w}), \qquad \text{by (6.10)}, \\
&= \Re 2\pi i v_\lambda(w,\overline{w}) + 2\pi\Im v_\lambda(w,\overline{w}), \qquad \text{by (6.18)}, \\
&= 0.
\end{aligned} \tag{7.20}
$$

Similarly for $B_{\lambda'}$.

For any self-symmetric circuit B_r, the separate terms of G, as expressed by (7.19), are readily verified, with the use of (7.13) and (6.9), to have the period zero. This results from the vanishing of the factor

$$
\begin{aligned}
(7.21)\qquad \Re 2i(\tau_{\alpha r} - \tau_{\alpha' r}) &= \Re 2i(\tau_{\alpha r} + \bar{\tau}_{\alpha r}) \\
&= \Re\ (\text{pure imag.}) = 0.
\end{aligned}
$$

Also, for every circuit A_j, the period of G vanishes. For, by property (b), that is true for Ω, and the complementary terms contribute the period

$$
(7.22)\qquad -T_{\alpha\beta}\Re 2i(\delta_j^\alpha - \delta_j^{\alpha'})\Im v_\beta(w,\overline{w}) = 0.
$$

Therefore the *uniformity* of G on R is established.

That properties (II) and (III) of G remain preserved is clear, since the complementary terms in (7.17) are everywhere regular analytic functions.

As to property (IV), we have abbreviating (7.17) to

$$
(7.23)\qquad G(z,w) = \Re S(z;w,\overline{w}),
$$

that

$$
(7.24)\qquad S(\bar{z};w,\overline{w}) = -\overline{S(z;w,\overline{w})}.
$$

This is easily seen from (7.14), (6.8) and (7.18). By taking real parts of (7.24), we get the desired property (IV); and the characterization of the formula (7.19) as Green's function is complete.

7. The functions ζ_j, $\wp_{jk}$

In the theory of elliptic functions, we have the fundamental Weierstrass function $\sigma(u)$, and in terms of this the other two Weierstrass functions $\zeta(u)$ and $\wp(u)$ are defined:

$$(7.25) \qquad \zeta(u) = \frac{d \log \sigma(u)}{du}, \qquad \wp(u) = -\frac{d\zeta(u)}{du} = -\frac{d^2 \log \sigma(u)}{du^2}.$$

$\sigma(u)$ with the periods 1, τ is essentially equivalent to the Jacobi θ-function $\theta(u, \tau)$ with the characteristic $1/2, 1/2$. (In this case $p = 1$, and the θ-series has only one summation index n, independent variable u, and period τ.) In fact,

$$(7.26) \qquad \sigma(u) = Ae^{Bu^2}\theta(u),$$

where A, B are constants as to u, depending only on τ:

$$(7.26') \qquad A = \frac{1}{\theta'(0;\tau)}, \qquad B = -\frac{\theta'(\frac{1}{2};\tau)}{\theta(\frac{1}{2};\tau)}.$$

We may therefore, with only slight change of the Weierstrass definitions, write

$$(7.27) \qquad \zeta(u) = \frac{d \log \theta(u)}{du}, \qquad \wp(u) = -\frac{d\zeta(u)}{du} = -\frac{d^2 \log \theta(u)}{du^2}.$$

These definitions generalize of themselves to multiple θ-series (Weierstrass, Riemann), giving systems of functions instead of a single function:

$$(7.28) \qquad \zeta_j(u) = \frac{\partial \log \theta(u)}{\partial u_j}, \qquad \wp_{jk}(u) = -\frac{\partial \zeta_j(u)}{\partial u_k} = -\frac{\partial^2 \log \theta(u)}{\partial u_j \partial u_k}.$$

We shall also consider

$$(7.29) \qquad \wp_{jkl}(u) = \frac{\partial \wp_{jk}(u)}{\partial u_l} = -\frac{\partial^3 \log \theta(u)}{\partial u_j \partial u_k \partial u_l}.$$

The periods τ_{jk} and the characteristic $\frac{1}{2}\beta$, $\frac{1}{2}\alpha$ are left implicit.

Since $\theta(u)$ with odd characteristic is an odd function, the same property follows for ζ_j :

$$(7.30) \qquad \zeta_j(-u) = -\zeta_j(u).$$

Also $\wp_{jk}$ is evidently an even function,

$$(7.31) \qquad \wp_{jk}(-u) = \wp_{jk}(u),$$

and is symmetric in its two indices. Finally, $\wp_{jkl}(u)$ is odd, and symmetric in its three indices.

We have, by logarithmic differentiation of (7.7,8), the following period properties of the functions ζ_j :

$$(7.32) \qquad \zeta_j(u_k + \delta_m^k) = \zeta_j(u),$$

$$(7.33) \qquad \zeta_j(u_k + \tau_{km}) = \zeta_j(u) - 2\pi i \delta_m^j.$$

From this, it follows by differentiation as to u_k that $\wp_{jk}(u)$ is $2p$-fold periodic in the exact sense:

$$\wp_{jk}(u_l + \delta_m^l) = \wp_{jk}(u), \tag{7.34}$$

$$\wp_{jk}(u_l + \tau_{lm}) = \wp_{jk}(u). \tag{7.35}$$

If, with reference to a Riemann surface R, we substitute $u_j = v_j(z, w)$ and for τ_{jk} the periods of v_j, then this periodicity of $\wp_{jk}(u)$ implies that $\wp_{jk}(v(z, w); \tau)$ is a uniform function of z on R, and evidently has only polar singularities, corresponding to the zeros of the denominator θ^2. In other words, $\wp_{jk}(v(z, w); \tau)$ is a rational function of z on R. Evidently, the same considerations apply to $\wp_{jk}(v(z, w); \tau)$.

On the other hand, according to (7.33), the function $\zeta_j(v(z, w); \tau)$ has the period $-2\pi i\delta_k^j$ with respect to the circuit B_k, although its periods as to the circuits A_k vanish, by (7.32). It has only polar singularities, corresponding to the zeros of $\theta(v(z, w); \tau)$, which are p in number, one located at w. These properties characterize $\zeta_j(v(z, w); \tau)$ as an integral of the second kind.

We define also the *normal integral of the second kind*:

$$t(z, w) = -\frac{\partial \log \theta(v(z, w))}{\partial w} = \zeta_j(v(z, w))v_j'(w) \tag{7.36}$$

(Riemann's notation). Evidently the only singularity of $t(z, w)$ is a pole of first order at $z = w$, with residue equal to unity. By the period properties of $\zeta_j(v(z, w))$, the periods of $t(z, w)$ as to the circuits A_k all vanish, while the period of $t(z, w)$ as to B_k is $-2\pi iv_k'(w)$.

§8. Solution of the Dirichlet Problem for a General Riemann Surface

1

Since a minimal surface is definable by harmonic functions, the solution of the Dirichlet problem is a prerequisite for that of the Plateau problem.

Let R denote a symmetric Riemann surface of the first kind, whose transition curves C bound the semi-surface R'. We may consider the Dirichlet problem for R': given on the boundaries C a continuous function $g(z)$, to find a function $H(u, v)$ uniform and continuous on R', harmonic in its interior, and coinciding with $g(z)$ on C.

In terms of Green's function for R, the solution of the Dirichlet problem is given by the classic formula

$$H(u, v) = -\frac{1}{2\pi}\int_C g(z)\frac{\partial G(z, w)}{\partial n}\,ds \qquad (w = u + iv). \tag{8.1}$$

Here $\partial/\partial n$ denotes differentiation in the direction of the interior normal at z.

Every harmonic function can be expressed as the real part of an analytic function of a complex variable:

$$H(u, v) = \Re F(w), \tag{8.2}$$

and an equivalent form of solution of the Dirichlet problem would be to find $F(w)$ explicitly. This we shall now do, following, with some additions, the exposition in our paper on the two-contour case.

It is to be remarked that while $H(u, v)$ must be uniform, $F(w)$ is not necessarily uniform, but only $\Re F(w)$; the function $F(w)$ itself may have pure imaginary periods with respect to the circuits of R'.

2

We first write the expression (7.17) for Green's function in the form

$$G(z, w) = \Re S(z; w, \overline{w}) = \Re[S(z, w) - S(z, \overline{w})], \tag{8.3}$$

where (see (7.2))

$$S(z, w) = \log \theta(v(z, w)) - T_{\alpha\beta}[v_\alpha(z) - v_{\alpha'}(z)][v_\beta(w) - v_{\beta'}(w)]. \tag{8.4}$$

We shall also use the functions

$$Z(z, w) = \frac{\partial S(z, w)}{\partial z}, \qquad P(z, w) = \frac{\partial^2 S(z, w)}{\partial z \partial w}. \tag{8.5}$$

The notation S, Z, P is intended to recall the elliptic functions $\log \sigma$, ζ, $\wp$, to which the former are equivalent in the case $p = 1$. By (8.4),

$$Z(z, w) = \zeta_j(v(z, w))v_j'(z) - T_{\alpha\beta}[v_\alpha'(z) - v_{\alpha'}'(z)][v_\beta(w) - v_{\beta'}(w)], \tag{8.6}$$

$$P(z, w) = \wp_{jk}(v(z, w))v_j'(z)v_k'(w) - T_{\alpha\beta}[v_\alpha'(z) - v_{\alpha'}'(z)][v_\beta'(w) - v_{\beta'}'(w)]. \tag{8.7}$$

Let $G_1(z)$ denote the conjugate harmonic function to $G(z)$; then, by definition,

$$\frac{\partial G_1}{\partial s} = -\frac{\partial G}{\partial n}. \tag{8.8}$$

By substitution in (8.1),

$$H(u, v) = \frac{1}{2\pi}\int_C g(z)\,\frac{\partial G_1(z)}{\partial s}\,ds = \frac{1}{2\pi}\int_C g(z)\,dG_1(z). \tag{8.9}$$

We have, by reference to (8.3),

$$G(z) + iG_1(z) = S(z; w, \overline{w}); \tag{8.10}$$

and on C, $G(z) = 0$; therefore on C,

$$idG_1(z) = dS(z; w, \overline{w}). \tag{8.11}$$

Hence, by (8.9),

$$\begin{aligned} H(u, v) &= \frac{1}{2\pi i}\int_C g(z)\,dS(z; w; \overline{w}) \\ &= \frac{1}{2\pi i}\int_C g(z)\,dS(z, w) - \frac{1}{2\pi i}\int_C g(z)\,dS(z, \overline{w}) \\ &= \frac{1}{2\pi i}\int_C g(z)Z(z, w)\,dz - \frac{1}{2\pi i}\int_C g(z)Z(z, \overline{w})\,dz. \end{aligned} \tag{8.12}$$

This formula expresses the harmonic function H in terms of so-called minimal coördinates $w = u + iv$, $\overline{w} = u - iv$, and has the characteristic form for such functions:

$$H(u, v) = \varphi(w) - \psi(\overline{w}). \tag{8.13}$$

In the variables w, $\overline{w}$, the Cauchy-Riemann equations for conjugate functions H, K become, as is easily calculated,

$$\frac{\partial H}{\partial w} = \frac{\partial (iK)}{\partial w}, \qquad \frac{\partial H}{\partial \overline{w}} = -\frac{\partial (iK)}{\partial \overline{w}}. \tag{8.14}$$

Accordingly, by (8.10), the conjugate harmonic function K to H obeys the equations

$$\frac{\partial (iK)}{\partial w} = \varphi'(w), \qquad \frac{\partial (iK)}{\partial \overline{w}} = \psi'(\overline{w}). \tag{8.15}$$

Therefore, if we fix the arbitrary additive constant inherent in K by the stipulation that at any chosen point (u_0, v_0), we shall have

$$K(u_0, v_0) = 0, \tag{8.16}$$

the equations (8.15) give

$$iK = \varphi(w) + \psi(\overline{w}) - [\varphi(w_0) + \psi(\overline{w}_0)]. \tag{8.17}$$

Since K is defined by the integral

$$K = -\int_{w_0}^{w} \frac{\partial H}{\partial n}\, ds \tag{8.18}$$

along a path on R', it is in general *multiform*, having periods with respect to the various circuits of R. The function H, on the other hand, is by definition uniform. This could also be verified from the explicit formula (8.12), since

$$Z(z, w) - Z(z, \overline{w})$$

is a uniform function of both points z and u, v.

Writing, in accordance with (8.2),

$$H(u, v) + iK(u, v) = F(w), \tag{8.19}$$

we have by (8.13), (8.17),

$$F(w) = 2\varphi(w) - [\varphi(w_0) + \psi(w_0)]; \tag{8.20}$$

therefore, by reference to (8.12, 13),

$$F(w) = \frac{1}{\pi i}\int_C g(z)Z(z, w)\, dz - \frac{1}{2\pi i}\int_C g(z)[Z(z, w_0) + Z(z, \overline{w}_0)]\, dz. \tag{8.21}$$

This is the desired explicit formula for $F(w)$.

The second term in this formula is a constant, arranged, as stated in connection with (8.16), so that

$$\Im F(w_0) = 0. \tag{8.22}$$

We repeat that only $\Re F(w)$ is uniform, while $F(w)$ itself generally has pure imaginary periods with respect to the circuits of R'. The condition (8.22) is required to hold for one of the branches of $F(w)$.

3. The condition for a minimal surface

In the application to the Plateau problem, $g(z)$, $H(u, v)$, $F(w)$ are *vectors* with n components, that may be distinguished by means of an index i varying from 1 to n.[34]

The condition that the surface

$$x_i = H_i(u, v) = \Re F_i(w) \tag{8.23}$$

be minimal is

$$\sum_{i=1}^{n} F_i'^2(w) = 0. \tag{8.24}$$

By differentiation of (8.21) as to w, we have with the notation (8.5),

$$F_i'(w) = \frac{1}{\pi i} \int_C g_i(z) P(z, w)\, dz. \tag{8.25}$$

Therefore,

$$\sum_{i=1}^{n} F_i'^2(w) = -\frac{1}{\pi^2} \int_C \int_C \sum_{i=1}^{n} g_i(z) g_i(\zeta) \cdot P(z, w) P(\zeta, w)\, dz\, d\zeta. \tag{8.26}$$

Writing vectorially

$$[g(z) - g(\zeta)]^2 = \sum_{i=1}^{n} [g_i(z) - g_i(\zeta)]^2, \tag{8.27}$$

we may put this formula in the form

$$\sum_{i-1}^{n} F_i'^2(w) = \frac{1}{2\pi^2} \int_C \int_C [g(z) - g(\zeta)]^2 P(z, w) P(\zeta, w)\, dz\, d\zeta. \tag{8.28}$$

Each of the added terms has the factor

$$\int_C P(z, w) dz = [Z(z, w)]_C = 0, \tag{8.29}$$

where the indicated variation of the multiform function $Z(z, w)$ over C vanishes, because C bounds the semi-surface R' and is therefore equivalent to a zero circuit.

---

[34] i the index, and $i = \sqrt{-1}$, are easily told apart.

§9. The Functional $A(g, R)$

1

We have defined

$$A(g, R) = D(H) = \frac{1}{2}\int\int_{R'} \sum_{i=2}^{n} \left[\left(\frac{\partial H_i}{\partial u}\right)^2 + \left(\frac{\partial H_i}{\partial v}\right)^2\right] du\, dv, \tag{9.1}$$

where H denotes the harmonic vector function on R' with the boundary values g. In this section, we obtain an explicit formula for $A(g, R)$ as a double contour integral in terms of g and quantities depending on the Riemann surface R.

By Green's formula,

$$\int\int_{R'} \left(\frac{\partial P}{\partial u}\frac{\partial Q}{\partial u} + \frac{\partial P}{\partial v}\frac{\partial Q}{\partial v}\right) du\, dv = \int\int_{R'} P\left(\frac{\partial^2 Q}{\partial u^2} + \frac{\partial^2 Q}{\partial v^2}\right) du\, dv - \int_C P\frac{dQ}{dn}\, ds, \tag{9.2}$$

where d/dn refers to the interior sense of the normal.

If $P = Q = H$, a harmonic function, this becomes

$$\int\int_{R'} \left[\left(\frac{\partial H}{\partial u}\right)^2 + \left(\frac{\partial H}{\partial v}\right)^2\right] du\, dv = -\int_C H\frac{dH}{dn}\, ds. \tag{9.3}$$

However, the behavior of the partial derivatives of H as we approach the boundary C of R' is uncertain. For that reason, we first slightly contract R' to $\bar{R}'$, bounded by curves $\bar{C}$ interior to R' and arbitrarily close to C. Then we shall pass to the limit as $\bar{C}$ tends to C, $\bar{R}'$ to R'.

We start, then, with

$$\begin{aligned}\bar{A}(g, R) &= \frac{1}{2}\int\int_{\bar{R}'} \sum_{i=1}^{n} \left[\left(\frac{\partial H_i}{\partial u}\right)^2 + \left(\frac{\partial H_i}{\partial v}\right)^2\right] du\, dv \\ &= -\frac{1}{2}\int_{\bar{C}} \sum_{i=1}^{n} H_i\frac{dH_i}{dn}\, ds.\end{aligned} \tag{9.4}$$

Writing, as in (8.19),

$$H_i(u, v) + iK_i(u, v) = F_i(w), \tag{9.5}$$

we have by the definition (8.18) of conjugate harmonic function,

$$-\frac{dH_i}{dn}\, ds = dK_i = \frac{1}{i}[dF_i(w) - dH_i]. \tag{9.6}$$

Substituting in (9.4), we get

$$\bar{A}(g, R) = \frac{1}{2i}\int_{\bar{C}} \sum_{i=1}^{n} H_i F_i'(w)\, dw - \frac{1}{2i}\int_{\bar{C}} \sum_{i=1}^{n} H_i\, dH_i. \tag{9.7}$$

The last term vanishes, because its indefinite integral is the uniform function $\frac{1}{2}\sum_{i=1}^{n} H_i^2$, and the path of integration $\bar{C}$ consists of closed circuits. Thus,

$$\bar{A}(g, R) = \frac{1}{2i}\int_{\bar{C}} \sum_{i=1}^{n} H_i(u, v)F_i'(w)\, dw. \tag{9.8}$$

By substitution of (8.25), this becomes

$$\bar{A}(g, R) = -\frac{1}{2\pi}\int_C \int_{\bar{C}} \sum_{i=1}^{n} g_i(z)H_i(u, v)\cdot P(z, w)\, dz\, dw. \tag{9.9}$$

In order to pass to the limit, we first make the preliminary transformation:

$$\bar{A}(g, R) = \frac{1}{4\pi}\int_C \int_{\bar{C}} \sum_{i=1}^{n} [g_i(z) - H_i(u, v)]^2 P(z, w)\, dz\, dw, \tag{9.10}$$

where the added terms vanish by (8.29).

Here we may pass to the limit formally as $\bar{C} \to C$, where w on $\bar{C}$ tends to ζ on C, and $H_i(u, v) \to g_i(\zeta)$, so that we get

$$A(g, R) = \frac{1}{4\pi}\int_C \int_C [g(z) - g(\zeta)]^2 P(z, \zeta)\, dz\, d\zeta. \tag{9.11}$$

The function-theoretic justification for this passage to the limit under the sign depends on the boundedness of the ratio $P(z, w)$: $P(z, \zeta)$ as $w \to \zeta$ along a normal to C. The complete discussion has been given for the one-contour case,[35] and the extension to the general case presents no essential difficulties.

Although complex elements enter into the expression for the functional $A(g, R)$, the value of that expression must be positive real on account of the interpretation in terms of Dirichlet integrals. Indeed, each element of the double integral (9.11) must be positive real.

We observe that the integral (9.11) is improper, since for $z = \zeta$, $P(z, \zeta)$ has a singularity with principal term $1/(z - \zeta)^2$. This follows from the defining formulas (8.4, 5), because $\theta(v(z, \zeta))$ vanishes when $z = \zeta$. To define (9.11), we must integrate first with $|\, z - \zeta\,| \geqq \epsilon$, and then take the limit as $\epsilon \to 0$.

Substituting (8.7) in (9.11), we have *in extenso*

$$\begin{aligned} A(g, R) = {} & \frac{1}{4\pi}\int_C \int_C [g(z) - g(\zeta)]^2 \wp_{jk}(v(z, \zeta))\, dv_j(z) dv_k(\zeta) \\ & - \frac{1}{4\pi}\int_C \int_C [g(z) - g(\zeta)]^2 T_{\alpha\beta}[dv_\alpha(z) - dv_{\alpha'}(z)][dv_\beta(\zeta) - dv_{\beta'}(\zeta)]. \end{aligned} \tag{9.12}$$

[35] [5], pp. 309–310.

We shall have to relate this to the formula obtained by making the same substitution in (8.28):

$$\sum_{i=1}^{n} F_i'^2(w) = \frac{1}{2\pi^2}\int_c\int_c [g(z) - g(\zeta)]^2 \wp_{jk}(v(z,\zeta))\, dv_j(z) dv_k(\zeta)$$

$$- \frac{1}{\pi^2}\int_c\int_c [g(z) - g(\zeta)]^2 \wp_{jk}(v(z,w))\, dv_j(z) v_k'(w)$$

$$\cdot T_{\alpha\beta}[dv_\alpha(\zeta) - dv_{\alpha'}(\zeta)][v_\beta'(w) - v_{\beta'}'(w)] \tag{9.13}$$

$$+ \frac{1}{2\pi^2}\int_c\int_c [g(z) - g(\zeta)]^2 T_{\alpha\lambda} T_{\beta\mu}[dv_\alpha(z) - dv_{\alpha'}(z)]$$

$$[dv_\beta(\zeta) - dv_{\beta'}(\zeta)][v_\lambda'(w) - v_{\lambda'}'(w)][v_\mu'(w) - v_{\mu'}'(w)].$$

2. Principal and complementary terms

In the cases $r = 0, 1$ of self-symmetry, only the first terms are present in the last two formulas. Accordingly, we separate these into "principal" and "complementary" terms, writing

$$A(g, R) = \mathcal{P}[A(g, R)] + \mathcal{C}[A(g, R)], \tag{9.14}$$

$$\sum_{i=1}^{n} F_i'^2(w) = \mathcal{P}\left[\sum_{i=1}^{n} F_i'^2(w)\right] + \mathcal{C}\left[\sum_{i=1}^{n} F_i'^2(w)\right]. \tag{9.15}$$

The explicit values of $\mathcal{P}$ and $\mathcal{C}$ can be seen by reference to (9.12), (9.13).

§10. The Identity

1

In the fundamental case of a simply-connected minimal surface bounded by a single contour, let us take for R the z-plane or sphere ($p = 0$), and for R' the half-plane or hemisphere $\Im z > 0$. Then Green's function is

$$G(z, w) = \Re \log \frac{z - w}{z - \overline{w}}.$$

We have by comparison with the general formula (8.3),

$$S(z, w) = \log(z - w),$$

and by (8.5),

$$Z(z, w) = \frac{1}{z - w}, \qquad P(z, w) = \frac{1}{(z - w)^2}.$$

We easily verify the algebraic identity

$$\frac{1}{(z-w)^2} + \frac{1}{(z-\zeta)^2} + \frac{1}{(\zeta - w)^2} = \left[\frac{1}{z-w} - \frac{1}{z-\zeta} - \frac{1}{\zeta - w}\right]^2, \tag{10.1}$$

which may be expressed as follows:

$$(10.2)\quad P(z, w) + P(z, \zeta) + P(\zeta, w) = [Z(z, w) - Z(z, \zeta) - Z(\zeta, w)]^2.$$

2

Let R' denote the strip of the z-plane $0 \leqq \Re z \leqq \frac{\pi}{2}$. This corresponds to the case of two contours with a single point in common.[36] We have

$$G(z, w) = \log \frac{\sin(z - w)}{\sin(z + \overline{w})}.$$

By comparison with (8.3),

$$S(z, w) = \log \sin(z - w), \qquad Z(z, w) = \cot(z - w), \qquad P(z, w) = \csc^2(z - w).$$

We have the trigonometric identity

$$(10.3)\quad \csc^2(z - w) + \csc^2(z - \zeta) + \csc^2(\zeta - w) = [\cot(z - w) - \cot(z - \zeta) - \cot(\zeta - w)]^2,$$

which can be directly verified. In terms of Z and P, this can again be expressed in the form (10.2).

3

In the case of two contours with no point in common, and a doubly-connected minimal surface, the Riemann surface R is a torus ($p = 1$). The corresponding θ-function $\theta(u; \tau)$ with the characteristic $1/2, 1/2$ is essentially equivalent to the elliptic σ-function of Weierstrass (formula 7.26).

Let ζ, $\wp$ denote the other elliptic functions of Weierstrass, (7.25), as based on σ. Then these have the addition theorems:[37]

$$(10.4)\quad \wp(u_1 + u_2) + \wp(u_1) + \wp(u_2) = \frac{1}{4}\left\{\frac{\wp'(u_1) - \wp'(u_2)}{\wp(u_1) - \wp(u_2)}\right\}^2,$$

$$(10.5)\quad \zeta(u_1 + u_2) - \zeta(u_1) - \zeta(u_2) = \frac{1}{2}\,\frac{\wp'(u_1) - \wp'(u_2)}{\wp(u_1) - \wp(u_2)}.$$

These formulas combine to give

$$(10.6)\quad \wp(u_1 + u_2) + \wp(u_1) + \wp(u_2) = [\zeta(u_1 + u_2) - \zeta(u_1) - \zeta(u_2)]^2.$$

If instead we use a ζ, $\wp$ based on θ according to the formulas (7.27), this identity takes the slightly modified form

$$(10.7)\quad \wp(u_1 + u_2) + \wp(u_1) + \wp(u_2) = [\zeta(u_1 + u_2) - \zeta(u_1) - \zeta(u_2)]^2 - 6B,$$

where B, constant as to u_1, u_2, is defined by (7.26′).

---

[36] Cf. [8], where the strip $0 \leqq \Im z \leqq \pi$ was used, and accordingly hyperbolic functions appeared instead of trigonometric.

[37] Hurwitz-Courant, *Funktionentheorie*, 1925, p. 189.

On the torus there is just one normal integral of first kind $v(z)$. Let us write $u_1 = v(z, \zeta)$, $u_2 = v(\zeta, w)$; then $u_1 + u_2 = v(z, w)$, and (10.7) becomes[38]

$$\wp(v(z, w)) + \wp(v(z, \zeta)) + \wp(v(\zeta, w)) = [\zeta(v(z, w)) - \zeta(v(z, \zeta)) - \zeta(v(\zeta, w))]^2 - 6B. \tag{10.8}$$

The analogy with (10.1, 2, 3) is apparent.

4

The identities (10.1, 3, 8) played an important rôle in the author's method for solving those cases of the Plateau problem in connection with which they were cited. To deal with the general case, an identity was needed which would take the place of these for a Riemann surface of arbitrary genus p.

As far as I know, the required identity did not exist in the literature of Riemann surfaces or θ-functions. It was necessary to discover and prove it, so as to be able to make the appropriate application to the Plateau problem.

I found the requisite general identity, applying to a Riemann surface of any genus p, to be the following:[38a]

$$\begin{aligned}(\mathrm{I}) \quad & [\wp_{lm}(v(z, w)) + \wp_{lm}(v(z, \zeta)) + \wp_{lm}(v(\zeta, w))]v_l'(w)v_m'(w) = [\zeta_l(v(z, w)) \\ & - \zeta_l(v(z, \zeta)) - \zeta_l(v(\zeta, w))][\zeta_m(v(z, w)) - \zeta_m(v(z, \zeta)) - \zeta_m(v(\zeta, w))] \\ & v_l'(w)v_m'(w) + R(w),\end{aligned}$$

where $R(w)$ denotes a rational function of w alone, i.e. a function uniform on R and with only polar singularities.

It seems interesting that (I) should first appear as a by-product of the solution of the Plateau problem; its more elementary forms (10.1, 3, 8) were previously well-known. Perhaps this is because these preliminary forms can be stated in terms of independent variables z, ζ, w or u_1, u_2 without reference to a Riemann surface; whereas for $p > 1$, the formula (I) requires a Riemann surface to carry the p normal integrals v_l of the first kind. In the case $p = 1$, there is only one normal integral $v_1(w)$, and $v_l'(w)v_m'(w)$ reduces to the single factor $v_1'^2(w)$, while $R(w) = -6Bv_1'^2(w)$, so that the factor $v_1'^2(w)$ may be cancelled from both sides of the identity, giving the form (10.8).

5

Proceeding to the proof of (I), we write

$$\begin{aligned}\phi(z, \zeta, w) = & [\wp_{lm}(v(z, w)) + \wp_{lm}(v(z, \zeta)) + \wp_{lm}(v(\zeta, w))]v_l'(w)v_m'(w) \\ & - [\zeta_l(v(z, w)) - \zeta_l(v(z, \zeta)) - \zeta_l(v(\zeta, w))][\zeta_m(v(z, w)) - \zeta_m(v(z, \zeta)) \\ & - \zeta_m(v(\zeta, w))]v_l'(w)v_m'(w).\end{aligned} \tag{10.9}$$

[38] It should be easy to distinguish ζ the function from ζ the argument.

[38a] The rest of this section, concerned with the indentities (I) and (II), follows [2], pp. 116–121, where these identities were first stated and proved.

We have to show that $\phi(z, \zeta, w)$ is really a rational function of w alone.

First, consider ϕ as a function only of z on the Riemann surface R.

Then $\wp_{lm}(v(z, w))$, $\wp_{lm}(v(z, \zeta))$ are uniform functions of z, while $\zeta_l(v(z, w))$ and $\zeta_l(v(z, \zeta))$ have the period zero with respect to every circuit of the first system, and the period $-2\pi i\delta^l_m$, the same for both, with respect to any circuit B_m of the second system. It follows that ϕ is a *uniform* function of z on R. Evidently, too, it has at most polar singularities since these are the only ones manifested by $\wp_{lm}$ or ζ_l, ζ_m. Consequently, ϕ is a *rational* function of z on R.

By similar reasoning, ϕ is rational on R in the variables ζ and w. Accordingly, we may write

$$\phi(z, \zeta, w) = R(z, \zeta, w). \tag{10.10}$$

The point is now to show that $R(z, \zeta, w)$ really does not depend on z or ζ.

Using the subscripts l, m to denote partial differentiation of $\theta(u)$ as to u_l, u_m, we have by the definitions (7.28),

$$\zeta_l = \frac{\theta_l}{\theta}, \qquad \wp_{lm} = -\frac{\partial \zeta_l}{\partial u_m} = \frac{\theta_l\theta_m}{\theta^2} - \frac{\theta_{lm}}{\theta} = \zeta_l\zeta_m - \frac{\theta_{lm}}{\theta}. \tag{10.11}$$

Substituting in (10.9) and referring to (10.10), we get

$$\begin{aligned} R(z, \zeta, w) = \Bigg\{ &-\frac{\theta_{lm}(v(z, w))}{\theta(v(z, w))} - \frac{\theta_{lm}(v(z, \zeta))}{\theta(v(z, \zeta))} - \frac{\theta_{lm}(v(\zeta, w))}{\theta(v(\zeta, w))} \\ &+ 2\frac{\theta_l(v(z, w))\theta_m(v(z, \zeta))}{\theta(v(z, w))\theta(v(z, \zeta))} + 2\frac{\theta_l(v(z, w))\theta_m(v(\zeta, w))}{\theta(v(z, w))\theta(v(\zeta, w))} \\ &- 2\frac{\theta_l(v(z, \zeta))\theta_m(v(\zeta, w))}{\theta(v(z, \zeta))\theta(v(\zeta, w))} \Bigg\} v'_l(w)v'_m(w). \end{aligned} \tag{10.12}$$

We first consider $R(z, \zeta, w)$ as a function of z, writing $R(z)$.

The function $\theta(v(z, \zeta))$ vanishes to the first order at $z = \zeta$, and at exactly $p - 1$ other points $a_1, a_2, \cdots, a_{p-1}$, *which are independent*[39] *of* ζ—so that the function $\theta(v(z, w))$ vanishes at $z = w$ and the *same* $p - 1$ other points.

The function $R(z)$ is regular except perhaps at those places where $\theta(v(z, \zeta)) = 0$ or $\theta(v(z, w)) = 0$. These are the points $z = \zeta, w, a_1, a_2, \cdots, a_{p-1}$, which might perhaps be poles.

We now prove that, in fact, $z = \zeta$ *is not a pole*, and by similar reasoning, *neither is* $z = w$.

Evidently, $z = \zeta$ is at most a pole of first order, and its residue is found by putting $z = \zeta$ in the numerator and in the derivative of the denominator of each fraction in (10.12) whose denominator vanishes at $z = \zeta$. This gives for the residue:

$$-\frac{\theta_{lm}(0)v'_l(w)v'_m(w)}{\theta_k(0)v'_k(\zeta)} + 2\frac{\zeta_l(v(\zeta, w))\theta_m(0)v'_l(w)v'_m(w)}{\theta_k(0)v'_k(\zeta)} - 2\frac{\zeta_m(v(\zeta, w))\theta_l(0)v'_l(w)v'_m(w)}{\theta_k(0)v'_k(\zeta)}.$$

[39] [17], p. 262 (IV).

The last two terms mutually cancel, differing only by the interchange of the summation indices l, m. This leaves only the first term, which is zero, because the θ-function with odd characteristic is, by (7.10), an odd function of the variables u_l; hence its Maclaurin series contains only odd powers of u; consequently all partial derivatives of even order vanish at $u = 0$; in particular, $\theta_{lm}(0) = 0$ for all l, m. The denominator $\theta_k(0)v_k'(\zeta)$ cannot vanish, since ζ, taken in general position, is a simple root of $\theta(v(z, \zeta))$.

In sum, $z = \zeta$ is a pole of first order with residue equal to zero; that is, it is not a pole at all.

A similar argument applies, as stated, to $z = w$.

It remains to examine the places $a_1, a_2, \cdots, a_{p-1}$. A look at the denominator of the term

$$2\frac{\theta_l(v(z, w))\theta_m(v(z, \zeta))}{\theta(v(z, w))\theta(v(z, \zeta))}v_l'(w)v_m'(w) \tag{10.13}$$

in (10.12) suggests that these places might be poles of the second order; but we now show that, in fact, they are at most poles of the first order.

For, as remarked above, the function $\theta(v(z, w))$ vanishes for $z = a_1$ regardless of the value of w (to repeat: the zeros of this function other than w are independent of w). That is, the equation $\theta(v(a_1, w)) = 0$ *is an identity in* w. Therefore it may be differentiated as to w, which gives

$$\theta_l(v(a_1, w))v_l'(w) = 0. \tag{10.14}$$

It follows that the constant term vanishes in the expansion of the numerator of (10.13) in powers of $z - a_1$; so that this expansion begins with $z - a_1$ to the first power at least. Since the expansion of the denominator begins with $(z - a_1)^2$, we see that a_1 is at most a pole of first order.

The same argument applies to $a_2, a_3, \cdots, a_{p-1}$. Therefore the rational function $R(z)$ is regular, except perhaps for first order poles at the $p - 1$ points $a_1, a_2, \cdots, a_{p-1}$.

It is an important theorem, going back to Weierstrass and related to the theorem of Riemann-Roch, that a rational function on a Riemann surface of genus p *cannot have fewer than* $p + 1$ *poles in general position, unless the function reduces to a constant.*[40] Each pole is to be counted with its proper multiplicity. For elliptic functions ($p = 1$), this is the well-known theorem that no non-constant elliptic function can have the total order of its poles less than two.

Applying this general theorem to the present situation, we have the result that $R(z, \zeta, w)$ *is a constant as to* z, provided that not all the $(p - 1)$-rowed determinants in the following p by $p - 1$ matrix vanish:

$$\begin{vmatrix} v_1'(a_1) & v_2'(a_1) & \cdots & v_p'(a_1) \\ v_1'(a_2) & v_2'(a_2) & \cdots & v_p'(a_2) \\ \cdots\cdots & \cdots\cdots & \cdots & \cdots\cdots \\ v_1'(a_{p-1}) & v_2'(a_{q-1}) & \cdots & v_p'(a_{p-1}) \end{vmatrix}. \tag{10.15}$$

[40] [18], p. 517.

This is the exact meaning of the phrase "in general position" appearing in the statement of the Weierstrass theorem.

To suppose that all the $(p-1)$-rowed minors of this matrix vanish would specialize the Riemann surface R, reducing the number of independent moduli on which it depends below the general value of $3p-3$. Supposing, then, that R is non-specialized, we have the result that $R(z, \zeta, w)$ does not really depend on z.

But $R(z, \zeta, w)$ is evidently symmetric in z, ζ; therefore $R(z, \zeta, w)$ *is also independent of* ζ.

Consequently, $R(z, \zeta, w)$ *is really a function of* w *alone*, $R(w)$, which was to be proved; and we have the desired identity (I).

Since the two members of the identity are analytic functions of all the independent variables and parameters involved, the identity, holding for non-specialized Riemann surfaces, must hold for all.

6

Actually, we shall not use the identity (I) in the form given, but rather in the one which results by applying the operator $\partial^2/\partial z\partial\zeta$, which has, among other circumstances, the advantage of eliminating the undetermined function $R(w)$.

We substitute in (10.12) the partial differential equation (7.6) of the θ-function, divide by two, and rearrange the terms, getting:

$$\begin{aligned}(10.16)\qquad &\Big\{\zeta_l(v(z,\zeta))\zeta_m(v(\zeta,w)) - \zeta_l(v(z,w))\zeta_m(v(z,\zeta)) + 2\pi i\,\frac{\partial \log\theta(v(z,\zeta))}{\partial\tau_{lm}} \\ &\quad + 2\pi i\,\frac{\partial \log\theta(v(z,w))}{\partial\tau_{lm}} + 2\pi i\,\frac{\partial \log\theta(v(\zeta,w))}{\partial\tau_{lm}}\Big\} v_l'(w)v_m'(w) \\ &\qquad = \zeta_l(v(z,w))\zeta_m(v(\zeta,w))v_l'(w)v_m'(w) - \tfrac{1}{2}R(w).\end{aligned}$$

Applying to this the operator $\partial^2/\partial z\partial\zeta$, we obtain the identity

$$\begin{aligned}(\mathrm{II})\qquad &\Big\{\wp_{jkl}(v(z,\zeta))[-\zeta_m(v(z,w)) + \zeta_m(v(\zeta,w))] + \wp_{kl}(v(z,\zeta))\wp_{jm}(v(z,w)) \\ &\quad + \wp_{jm}(v(z,\zeta))\wp_{kl}(v(\zeta,w)) + 2\pi i\,\frac{\partial\wp_{jk}(v(z,\zeta))}{\partial\tau_{lm}}\Big\} v_j'(z)v_k'(\zeta)v_l'(w)v_m'(w) \\ &\qquad = \wp_{jl}(v(z,w))\wp_{km}(v(\zeta,w))v_j'(z)v_k'(\zeta)v_l'(w)v_m'(w).\end{aligned}$$

In the course of the calculation, we interchange summation indices freely, and observe also that

$$\frac{\partial^2}{\partial z\partial\zeta}\,\frac{\partial \log\theta(v(z,\zeta);\tau)}{\partial\tau_{lm}} = \frac{\partial}{\partial\tau_{lm}}\,\frac{\partial^2 \log\theta(v(z,\zeta);\tau)}{\partial z\partial\zeta} = \frac{\partial\wp_{jk}(v(z,\zeta);\tau)}{\partial\tau_{lm}}\,v_j'(z)v_k'(\zeta).$$

§11. A Special Variation of the Representation g, R

1

Let w, $\bar{w}$ denote a pair of conjugate complex branch-points of the symmetric Riemann surface R, which we picture in the standard way as a many-sheeted surface over the complex x-plane or x-sphere. We recall that the conformal involution T of R may be represented as a reflection in the real axis $x \to \bar{x}$, combined with a substitution of period two on the various sheets. The transition curves C correspond to real branch-cuts. We may also represent R as an algebraic curve $P(x, y) = 0$ with real coefficients, regarded as a two-dimensional manifold in the space of the complex variables $x = x_1 + ix_2$, $y = y_1 + iy_2$. The transition curves C correspond then to the real branches in the xy-plane. The algebraic curve P is determined up to real birational transformation. This is equivalent to a conformal transformation of the corresponding Riemann surface R, with preservation of symmetric points. All branch-points may therefore be supposed of first order,[41] since this is the case with the Clebsch-Lüroth normal form,[42] to which the general form of Riemann surface is conformally equivalent.

Suppose that the form of the Riemann surface is varied by movement of the branch-points w, $\bar{w}$, all other branch-points remaining fixed. We suppose that the displacements of w, $\bar{w}$ are conjugate complex, so that R remains symmetric:

$$\delta w = \epsilon = \epsilon_1 + i\epsilon_2, \qquad \delta\bar{w} = \bar{\epsilon} = \epsilon_1 - i\epsilon_2. \tag{11.1}$$

Let the varied surface R_ϵ be represented on R so that points belonging to the same value of x in the same sheets correspond. Since the real branch-points have not been affected, it is evident that the transition curves C correspond to one another in a one-one continuous manner. In terms of the representation of C as the real branches of the algebraic curve corresponding to R, the variation may be pictured as in the following figure. The original curve is drawn full, the varied one dotted, and the two set into one-one continuous correspondence by associating points with the same abscissa x.

Evidently, if a set of contours (Γ) is represented as topological image of the transition curves (C) of R, the variation just described gives a topological representation of (Γ) on the transition curves (C_ϵ) of the varied surface R_ϵ. We may therefore consider this variation as applying to the representation (g, R) of the given contours.

The fact that the abscissa x of each real branch-point b_1, b_2, b_3, b_4 remains invariant is obviously essential to the one-one character of the correspondence between real branches. Otherwise, we would have a figure like the following, where the one-one character breaks down in the vicinity of the branch-point

---

[41] Like that pertaining to $\sqrt{z}$.

[42] [18], pp. 440–444.

whose abscissa x has been varied. In fact, the one-one continuous, and indeed conformal, nature of the correspondence by equal values of x between the Riemann surfaces R_ϵ and R, applies everywhere except in the vicinity of the moving branch-points w, $\overline{w}$.

2

Let us determine the effect of the preceding variation on the normal integrals v of the first kind.

The characteristic property of the integrals v is that, although having periods with respect to the circuits of R, they can be prolonged indefinitely on R without coming to any singularity, i.e. any branch of $v_j(z)$ can be expressed in the vicinity of any place on R as a convergent power series in the local uniformizing variable t. This will be $t = z - a$ for a finite place a not a branch-point; $t = 1/z$ for

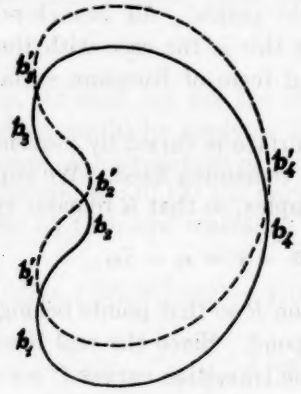


FIG. 19

FIG. 20

$a = \infty$; $t = \sqrt{(z - w)}$, $t = \sqrt{(z - \overline{w})}$ for the moving branch-points w, $\overline{w}$; and $t = \sqrt{(z - w')}$ for a fixed branch-point $w' \neq w$ or $\overline{w}$.

Thus, about the place a we have

$$v_j(z) = v_j(0) + v_j'(0)t + \cdots,$$

the argument 0 referring to $t = 0$. The coefficients are functions of w, $\overline{w}$. When these vary, we have according to (11.1),

$$(11.2) \qquad \delta v_j(z) = \left\{\frac{\partial v_j(0)}{\partial w}\epsilon + \frac{\partial v_j(0)}{\partial \overline{w}}\bar{\epsilon}\right\} + \left\{\frac{\partial v_j'(0)}{\partial w}\epsilon + \frac{\partial v_i'(0)}{\partial \overline{w}}\bar{\epsilon}\right\}t + \cdots$$

(t does not depend on w, $\overline{w}$)—which is regular about the place a.

About ∞, or about any branch-point $w' \neq w$ or $\overline{w}$, the same applies, for t does not depend on w, $\overline{w}$.

But about the moving branch-point w, we have $t = \sqrt{(z - w)}$, so that

$$(11.3) \qquad \frac{\partial t}{\partial w} = -\frac{1}{2\sqrt{(z - w)}} = -\frac{1}{2t}.$$

Therefore about this place we have the expansion

$$\delta v_j(z) = -\frac{\epsilon}{2} v_j'(0)\cdot\frac{1}{t} + \cdots, \tag{11.4}$$

the dots denoting a convergent power series in t.

Similarly, about $\overline{w}$ we have

$$\delta v_j(z) = -\frac{\bar{\epsilon}}{2} v_j'(0)\cdot\frac{1}{\bar{t}} + \cdots. \tag{11.5}$$

In sum, the function $\delta v_j(z)$ is regular everywhere on the Riemann surface R, except at w and $\overline{w}$, where it has first order poles with the respective residues

$$-\frac{\epsilon}{2} v_j'(w), \qquad -\frac{\bar{\epsilon}}{2} v_j'(\overline{w}),$$

the differentiations being with respect to the local uniformizing variable. Therefore $\delta v_j(z)$ is an integral of the second kind. Its periods with respect to the first system of circuits A_k are all zero, since those of $v_j(z)$ are constantly equal to the Kronecker symbols δ_k^j during the variation.

The normal integrals $t(z, w)$, $t(z, \overline{w})$ of the second kind with a unique pole at w, $\overline{w}$, respectively, have there the residue unity and, by definition, have also zero periods with respect to the first system of circuits. Hence the linear combination

$$-\frac{\epsilon}{2} v_j'(w)t(z, w) - \frac{\bar{\epsilon}}{2} v_j'(\overline{w})t(z, \overline{w})$$

is an integral of the second kind which has the same poles, with the same residues, and the same first system of periods as $\delta v_j(z)$. This identifies the two to within an additive constant, which may be adjusted to be zero; so that we have finally

$$\delta v_j(z) = -\frac{\epsilon}{2} v_j'(w)t(z, w) - \frac{\bar{\epsilon}}{2} v_j'(\overline{w})t(z, \overline{w}). \tag{11.6}$$

3

Now, our supposition that the conjugate points w, $\overline{w}$ were branch-points is not really a restriction, because, by real birational transformation, any two given conjugate points w, $\overline{w}$ can be made branch-points of R. This means that we can find a symmetric Riemann surface R_1, conformally equivalent to R with preservation of symmetric points, and such that any given symmetric points w, $\overline{w}$ of R are transformed into branch-points w_1, $\overline{w}_1$ of R_1.

This is easily seen geometrically as follows. Let the tangent lines t, $\bar{t}$ be drawn at w, $\overline{w}$ to the algebraic curve $\mathfrak{C}$ which corresponds to R; then t, $\bar{t}$ are conjugate imaginary straight lines and intersect in a real point O. Let the plane containing the curve $\mathfrak{C}$ be subjected to a real projective transformation that carries O into the infinite point of the y-axis; then the tangents at t, $\bar{t}$ become lines parallel to the y-axis, and their points of contact w, $\overline{w}$ become

branch-points w_1, $\overline{w}_1$. This projection of $\mathfrak{C}$ is a particular kind of real birational transformation.

It follows that in formula (11.6) we may suppose that w, $\overline{w}$ are *any* two conjugate points of R. The corresponding variation is PVP^{-1}, where P is the specified projection, and V the displacement of w_1, $\overline{w}_1$. Under PVP^{-1}, of course, *all* the branch-points of R vary in a special way.

4

In the variation (11.6) let us first allow ϵ, $\bar{\epsilon}$ to be two equal real displacements: $\epsilon = \epsilon_1$, $\bar{\epsilon} = \epsilon_1$; then

$$\delta_1 v_j(z) = \epsilon_1 \{-\tfrac{1}{2} v_j'(w) t(z, w) - \tfrac{1}{2} v_j'(\overline{w}) t(z, \overline{w})\}. \tag{11.7}$$

Then, let ϵ, $\bar{\epsilon}$ be two conjugate pure imaginary displacements: $\epsilon = i\epsilon_2$, $\bar{\epsilon} = -i\epsilon_2$; we get

$$\delta_2 v_j(z) = \epsilon_2 \left\{ -\frac{i}{2} v_j'(w) t(z, w) + \frac{i}{2} v_j'(\overline{w}) t(z, \overline{w}) \right\}. \tag{11.8}$$

Hereafter, for greater convenience, we shall use the symbol δ to denote a *derivative* instead of a differential:

$$\delta_1 v_j(z) = \left. \frac{\partial v_j(z, \epsilon_1)}{\partial \epsilon_1} \right|_{\epsilon_1 = 0}, \qquad \delta_2 v_j(z) = \left. \frac{\partial v_j(z, \epsilon_2)}{\partial \epsilon_2} \right|_{\epsilon_2 = 0};$$

and so we rewrite (11.7, 8):

$$\delta_1 v_j(z) = -\tfrac{1}{2} v_j'(w) t(z, w) - \tfrac{1}{2} v_j'(\overline{w}) t(z, \overline{w}), \tag{11.9}$$

$$\delta_2 v_j(z) = -\frac{i}{2} v_j'(w) t(z, w) + \frac{i}{2} v_j'(\overline{w}) t(z, \overline{w}). \tag{11.10}$$

5

Suppose that these variations δ_1, δ_2 of $v_j(z)$ are applied to formula (9.11) for $A(g, R)$. Then we get $\delta_1 A(g, R)$, $\delta_2 A(g, R)$, which may be calculated by performing the variations δ_1, δ_2 under the sign of integration on the factor which multiplies $[g(z) - g(\zeta)]^2$. The validity of this operation is easily established, with full account taken of the improper nature of the integral $A(g, R)$, by use of the assumption that $A(g, R)$ is finite. Indeed, $A(g, R)$ becomes an analytic function $A_1(\epsilon_1)$, $A_2(\epsilon_2)$ of ϵ_1, ϵ_2 respectively, whose power series expansion can be calculated formally.[42a]

We have therefore as variational conditions on $A(g, R)$:

$$\delta_1 A(g, R) = A_1'(0) = 0, \qquad \delta_2 A(g, R) = A_2'(0) = 0. \tag{11.11}$$

[42a] Cf. [5], pp. 289–297, where the sufficiently typical one-contour case is treated in detail.

We may combine these by forming

$$A_1'(0) - iA_2'(0) = 0. \tag{11.12}$$

It thus becomes only a question of how *formally* to calculate $A_1'(0) - iA_2'(0)$. We could do this literally by calculating separately the real and imaginary parts. But the calculations can be rendered considerably shorter and neater by the following device.

Evidently, $A(g, R)$ as calculated formally depends in a linear way on $\delta v_j(z)$. Hence, if $\delta_1 v_j(z)$ gives $A_1'(0)$, and $\delta_2 v_j(z)$ gives $A_2'(0)$, we may obtain $A_1'(0) - iA_2'(0)$ by applying *formally*[43]

$$\delta v_j(z) = \delta_1 v_j(z) - i\delta_2 v_j(z). \tag{11.13}$$

By (11.9, 10), we have simply

$$\delta v_j(z) = -v_j'(w)t(z, w). \tag{11.14}$$

Accordingly, we have for the corresponding formal variation of $A(g, R)$:

$$\delta A(g, R) = A_1'(0) - iA_2'(0), \tag{11.15}$$

the vanishing of which expresses the variational condition (11.12).

In (11.14), we may substitute for $t(z, w)$ its expression (7.36), getting

$$\delta v_j(z) = -v_j'(w)v_k'(w)\zeta_k(v(z, w)). \tag{11.16}$$

In this, allow z to describe the circuit B_m ; then since, by (7.33), the period of $\zeta_k(v(z, w))$ as to B_m is $-2\pi i\delta_m^k$, we obtain, with change of indices,

$$\delta\tau_{jk} = 2\pi i v_j'(w)v_k'(w). \tag{11.17}$$

§12. The First Variation of $A(g, R)$

1. Self-symmetric case $r = 0, 1$

In applying these variations to $A(g, R)$, let us consider first the self-symmetric case. Then (9.12) and (9.13) reduce to their principal terms; we have:

$$A(g, R) = \frac{1}{4\pi}\int_C\int_C [g(z) - g(\zeta)]^2 \varphi_{jk}(v(z, \zeta))\, dv_j(z)dv_k(\zeta). \tag{12.1}$$

$$\sum_{i=1}^{n} F_i'^2(w) = \frac{1}{2\pi^2}\int_C\int_C [g(z) - g(\zeta)]^2 \varphi_{jl}(v(z, w))\varphi_{km}(v(\zeta, w))dv_j(z)dv_k(\zeta)v_l'(w)v_m'(w). \tag{12.2}$$

[43] Cf. [2], p. 112.

On the elements of the integral $A(g, R)$ the effect of the variation (11.16, 17) is as follows:

$$\begin{aligned}\delta\varphi_{jk}(v(z,\zeta)) &= \varphi_{jkl}(v(z,\zeta))\{\delta v_l(z) - \delta v_l(\zeta)\} + \frac{\partial\varphi_{jk}(v(z,\zeta))}{\partial\tau_{lm}}\,\delta\tau_{lm}\\ &= -\varphi_{jkl}(v(z,\zeta))v_l'(w)v_m'(w)\zeta_m(v(\zeta,w))\\ &\quad + \varphi_{jkl}(v(z,\zeta))v_l'(w)v_m'(w)\zeta_m(v(\zeta,w))\\ &\quad + 2\pi i v_l'(w)v_m'(w)\,\frac{\partial\varphi_{jk}(v(z,\zeta))}{\partial\tau_{lm}};\end{aligned} \tag{12.3}$$

$$\delta dv_j(z) = d\delta v_j(z) = v_j'(w)v_l'(w)\,\varphi_{lm}(v(z,w))v_m'(z)dz; \tag{12.4}$$

$$\delta dv_k(\zeta) = d\delta v_k(\zeta) = v_k'(w)v_l'(w)\,\varphi_{lm}(v(\zeta,w))v_m'(\zeta)d\zeta. \tag{12.5}$$

By the product rule of differentiation,

$$\begin{aligned}&\delta[\varphi_{jk}(v(z,\zeta))dv_j(z)dv_k(\zeta)]\\ &\quad = dv_j(z)dv_k(\zeta)\delta\varphi_{jk}(v(z,\zeta)) + \varphi_{jk}(v(z,\zeta))dv_k(\zeta)\delta dv_j(z)\\ &\quad\quad + \varphi_{jk}(v(z,\zeta))dv_j(z)\delta dv_k(\zeta).\end{aligned} \tag{12.6}$$

From this, by substitution of (12.3, 4, 5), with change of summation indices to simplify the resulting expression, we get

$$\begin{aligned}&\delta[\varphi_{jk}(v(z,\zeta))dv_j(z)dv_k(\zeta)]\\ &\quad = \Big\{\varphi_{jkl}(v(z,\zeta))[-\zeta_m(v(z,w)) + \zeta_m(v(\zeta,w))]\\ &\quad\quad + \varphi_{kl}(v(z,\zeta))\varphi_{jm}(v(z,w)) + \varphi_{jm}(v(z,\zeta))\varphi_{kl}(v(\zeta,w))\\ &\quad\quad + 2\pi i\,\frac{\partial\varphi_{jk}(v(z,\zeta))}{\partial\tau_{lm}}\Big\}v_j'(z)v_k'(\zeta)v_l'(w)v_m'(w)\,dzd\zeta.\end{aligned} \tag{12.7}$$

Here we have exactly the first member of the identity (II) at the end of §10. Applying that identity, we obtain

$$\delta[\varphi_{jk}(v(z,\zeta))dv_j(z)dv_k(\zeta)] = \varphi_{jl}(v(z,w))\,\varphi_{km}(v(\zeta,w))dv_j(z)dv_k(\zeta)v_l'(w)v_m'(w). \tag{12.8}$$

By comparison with (12.2), we therefore have

$$\delta A(g, R) = \frac{\pi}{2}\sum_{i=1}^{n} F_i'^2(w). \tag{12.10}$$

Accordingly, the variational condition

$$\delta A(g, R) = 0 \tag{12.11}$$

is equivalent exactly to the condition for a minimal surface

$$\sum_{i=1}^{n} F_i'^2(w) = 0. \tag{12.12}$$

Since

$$\sum_{i=1}^{n} F_i'^2(w) = (E - G) - 2iF, \tag{12.13}$$

we may also express this result, with reference to (11.15), by resolution into real and imaginary parts, as follows:

$$\delta_1 A(g, R) = 0 \quad \text{implies} \quad E = G; \tag{12.14}$$

$$\delta_2 A(g, R) = 0 \quad \text{implies} \quad F = 0. \tag{12.15}$$

These are the conditions of conformality on the harmonic surface H, which characterize it as a minimal surface.

2. Alter-symmetric case, $r > 1$

In the alter-symmetric case, we must also take account of the complementary terms in $A(g, R)$ and $\sum_1^n F_i'^2(w)$.

In terms of the notation (9.14, 15), we may express the result (12.10) of the preceding article as follows:

$$\delta\mathcal{P}[A(g, R)] = \frac{\pi}{2}\,\mathcal{P}\left[\sum_{i=1}^{n} F_i'^2(w)\right]. \tag{12.16}$$

It remains then only to show that also

$$\delta\mathcal{C}[A(g, R)] = \frac{\pi}{2}\,\mathcal{C}\left[\sum_{i=1}^{n} F_i'^2(w)\right]. \tag{12.17}$$

This is a matter of straightforward calculation as follows.[43a]

We have, considering the essential factor of $\mathcal{C}[A(g, R)]$:

$$\begin{aligned}
&\delta\{T_{\alpha\beta}[dv_\alpha(z) - dv_{\alpha'}(z)][dv_\beta(\zeta) - dv_{\beta'}(\zeta)]\} \\
&\quad = T_{\alpha\beta}[d\delta v_\alpha(z) - d\delta v_{\alpha'}(z)][dv_\beta(\zeta) - dv_{\beta'}(\zeta)] \\
&\qquad + T_{\alpha\beta}[dv_\alpha(z) - dv_{\alpha'}(z)][d\delta v_\beta(\zeta) - d\delta v_{\beta'}(\zeta)] \\
&\qquad\quad + \delta T_{\alpha\beta}[dv_\alpha(z) - dv_{\alpha'}(z)][dv_\beta(\zeta) - dv_{\beta'}(\zeta)].
\end{aligned} \tag{12.18}$$

By (11.16),

$$\delta v_\alpha(z) - \delta v_{\alpha'}(z) = -[v'_\alpha(w) - v'_{\alpha'}(w)]\zeta_k(v(z, w))v'_k(w). \tag{12.19}$$

Therefore,

$$d\delta v_\alpha(z) - d\delta v_{\alpha'}(z) = [v'_\alpha(w) - v'_{\alpha'}(w)]\,\wp_{jk}(v(z, w))dv_j(z)v'_k(w). \tag{12.20}$$

Similarly,

$$d\delta v_\beta(\zeta) - d\delta v_{\beta'}(\zeta) = [v'_\beta(w) - v'_{\beta'}(w)]\,\wp_{jk}(v(\zeta, w))dv_j(\zeta)v'_k(w). \tag{12.21}$$

[43a] Cf. the preliminary note to the present paper, [31], particularly §11.

Repeating here the relation (6.18):

$$T_{\alpha\lambda}t_{\lambda\mu} = \delta_\mu^\alpha = \begin{cases} 1 & \text{if } \alpha = \mu, \\ 0 & \text{if } \alpha \neq \mu, \end{cases} \tag{12.22}$$

and applying to it the operator δ, we get

$$t_{\lambda\mu}\delta T_{\alpha\lambda} + T_{\alpha\lambda}\delta t_{\lambda\mu} = 0. \tag{12.23}$$

Multiply this by $T_{\beta\mu}$ and sum as to μ; then with use of (12.22),

$$\delta T_{\alpha\beta} + T_{\alpha\lambda}T_{\beta\mu}\delta t_{\lambda\mu} = 0. \tag{12.24}$$

Repeating (6.11),

$$t_{\alpha\beta} = \frac{1}{2\pi i}(\tau_{\alpha\beta} - \tau_{\alpha\beta'} - \tau_{\alpha'\beta} + \tau_{\alpha'\beta'}), \tag{12.25}$$

we have, by (11.17),

$$\delta t_{\alpha\beta} = [v'_\alpha(w) - v'_{\alpha'}(w)][v'_\beta(w) - v'_{\beta'}(w)]. \tag{12.26}$$

Substituting this in (12.24), we find, with change of indices,

$$\delta T_{\alpha\beta} = -T_{\alpha\lambda}T_{\beta\mu}[v'_\lambda(w) - v'_{\lambda'}(w)][v'_\mu(w) - v'_{\mu'}(w)]. \tag{12.27}$$

We may now make the substitutions (12.20, 21, 27) in (12.18), and get

$$\begin{aligned} &\delta\{T_{\alpha\beta}[dv_\alpha(z) - dv_{\alpha'}(z)][dv_\beta(\zeta) - dv_{\beta'}(\zeta)]\} \\ &= T_{\alpha\beta}[v'_\alpha(w) - v'_{\alpha'}(w)][dv_\beta(\zeta) - dv_{\beta'}(\zeta)]\,\wp_{jk}(v(z, w))dv_j(z)v'_k(w) \\ &+ T_{\alpha\beta}[v'_\beta(w) - v'_{\beta'}(w)][dv_\alpha(z) - dv_{\alpha'}(z)]\,\wp_{jk}(v(\zeta, w))dv_j(\zeta)v'_k(w) \\ &- T_{\alpha\lambda}T_{\beta\mu}[dv_\alpha(z) - dv_{\alpha'}(z)][dv_\beta(\zeta) - dv_{\beta'}(\zeta)][v'_\lambda(w) - v'_{\lambda'}(w)][v'_\mu(w) - v'_{\mu'}(w)]. \end{aligned} \tag{12.28}$$

If we substitute this in $\delta^{\mathcal{C}}[A(g, R)]$, interchange the indices α, β in the first term, the integration variables z, ζ in the second, and compare with (9.13), we obtain the desired relation (12.17).

3

By addition of (12.16) and (12.17), we have therefore in the general case

$$\delta A(g, R) = \frac{\pi}{2}\sum_{i=1}^{n} F_i'^2(w). \tag{12.29}$$

Accordingly, if we can prove the *attainment* of the minimum of $A(g, R)$, the existence of the required minimal surface will follow immediately.

§13. Representations (g, R), Proper and Improper as an L-set

1

We consider (g, R) as a composite mathematical element, consisting of a symmetric Riemann surface R of proper or reduced type (r, r'') together with a

topological or parametric representation g of the given contours (Γ) on the transition curves (C) of R. We recall that one of the manners in which R can become improper is for its transition curves (C) to have common points when the contours (Γ) have none.

The parametric representation of $\Gamma_1, \Gamma_2, \cdots, \Gamma_k$ may be pictured as a monotonic graph relating each Γ_j with the corresponding C_j, so that we have k such graphs, of which those represented in the following figure may be considered typical.

In the square $\Gamma_j C_j$, opposite points on parallel sides are regarded as identical, so that $\Gamma_j C_j$ is really a torus. Any topological correspondence between (Γ) and (C) is represented by a cyclically monotonic graph on each of the k tori $\Gamma_j C_j$. This means one such as $E_1E_2E_3E_4$, where we take into account the cyclic nature of the torus ($E_2 = E_3$, $E_4 = E_1$). We are thus led to include also improperly monotonic graphs, such as $A_1B_1B_2B_3B_4A_3$, with horizontal and

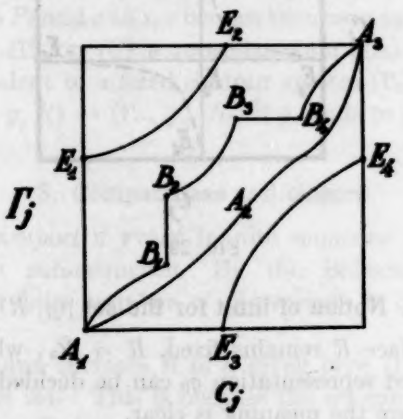


FIG. 21

vertical segments, the number of which is, by a known theorem, at most denumerably infinite. As extreme case, we have a completely degenerate graph, such as $E_1E_2E_3(=E_4)E_2E_5(=E_1)$ in fig. 22, consisting of a vertical and a horizontal transversal of the square, interpreted cyclically. These must be included because obtainable as limits of properly monotonic graphs, as the dotted curve in the figure illustrates.

With reference to these diagrams, we may distinguish then the following types of improper parametric representations g.

(a) *Continuously improper representations*, which cause a point of Γ_j to correspond to a partial arc of C_j, e.g. the arc B_3B_4 in fig. 21. Then g is still continuous, but does not establish a one-one correspondence between (Γ) and (C).

(b) *Discontinuously improper representations*—which cause a partial arc of a contour Γ_j to correspond to a point of C_j, e.g. the arc B_1B_2 in fig. 21. Then g is

discontinuous at the point of C_j, having there two unequal one-sided limits, namely, the *distinct* end-points of the corresponding arc of Γ_j.[43b]

(c) *Degenerate representations*, where the whole of Γ_j corresponds to a point of C_j, and the whole of C_j to a point of Γ_j. For a given contour Γ_j, the totality of degenerate representations corresponds to the choice of the two distinguished points on Γ_j and C_j respectively.

If we combine the diagram of fig. 21 with a visualization of all proper and reduced Riemann surfaces R of type (r, r''), we obtain a clear picture of the totality of representations $[(g, R)]$, proper and improper, of a given contour system (Γ).

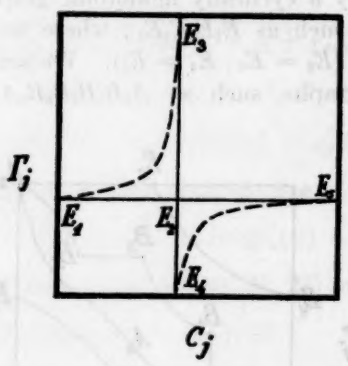


FIG. 22

2. Notion of limit for the set $[(g, R)]$

If the Riemann surface R remains fixed, $R = R_0$, while g varies, then the approach of g to a fixed representation g_0 can be decided by reference to the k monotonic graphs, where the meaning is clear.

If R is variable, then the notion of R approaching to a limit R_0 is easily defined. Whether R is represented as (1) a many-sheeted surface over the z-plane, (2) a plane region bounded by circles, or (3) the Riemann surface of an algebraic curve $P(x, y) = 0$, the meaning of the approach of R to R_0 is evident, being respectively as follows: (1) the branch-points of R tend to those of R_0, (2) the circles defining R tend to those of R_0, (3) the coefficients of $P(x, y)$ tend to those of $P_0(x, y)$. In the approach of R to R_0, we may suppose that their respective points z, z_0 are in one-one continuous correspondence so that $z \to z_0$ uniformly. In this correspondence, symmetry of point-pairs is supposed to be preserved, and therefore the transition curves (C) and (C_0) correspond.

The meaning of (g, R) tending to (g_0, R_0) may now be defined as follows. First, R must tend to R_0. Then the representation $g(z)$ of (Γ) on (C) can be transferred to a representation $\bar{g}(z_0)$ on (C_0) by the correspondence between z and z_0 just referred to. Since R_0 is fixed, the approach of $g(z_0)$ to $g_0(z_0)$ can be

[43b] See footnote 61a.

interpreted as before, by reference to the monotonic graph. This gives the meaning of $(g, R) \rightarrow (g_0, R_0)$.

The idea of *convergence* or *limit* being thus defined the system of all representations (g, R) of a given set of contours is an *L-set* in the sense of Fréchet [26]. An *L-set* is a class of elements in which the notion of limit has been defined in accordance with the following two postulates: (a) any partial sequence of a convergent sequence is also convergent and has the same limit as the original sequence; (b) a sequence which consists of indefinite repetition of the same element is convergent and has this element for limit.

For some purposes, it will be desirable to have a notion of limit also for the case where the contours (Γ) themselves vary: $(\Gamma, g, R) \rightarrow (\Gamma_0, g_0, R_0)$. First, we require (Γ) to tend to (Γ_0) in the sense of Fréchet.[44] This implies that the points P, P_0 of (Γ), (Γ_0) respectively can be coördinated so that $P \rightarrow P_0$ uniformly. Second, R must tend to R_0. This implies a correspondence z to z_0 between the transition curves (C) and (C_0) such that $z \rightarrow z_0$ uniformly. By the correspondence of P to P_0 and z to z_0, we can then associate with any parametric representation $\bar{g}(z)$ of (Γ) on (C) a representation $\bar{g}(z_0)$ of (Γ_0) on (C_0). The situation is now equivalent to a fixed contour system (Γ_0) and a fixed R_0. We shall then say that $(\Gamma, g, R) \rightarrow (\Gamma_0, g_0, R_0)$ if $\bar{g}$ tends to g_0 in the interpretation by monotonic graphs.

3. Compactness and closure

An L-set is called *compact* if every infinite sequence of elements of the set contains a convergent sub-sequence. By the Bolzano-Weierstrass theorem ordinary space of any finite number of dimensions is compact, if we include points at infinity.

All symmetric Riemann surfaces R of a given type (Γ, r, r'), proper or improper, form a compact set. This is because the set can be referred to a finite number of parameters, e.g., the position of the branch-points, or the centers and radii of the circles, or the coefficients of an algebraic equation of given degree, according to the mode of representation chosen for R. Essentially, these parameters are $3p - 3$ in number—the moduli.

Also the parametric representations proper or improper, of any given contour Γ_j form a compact set. For these have been referred to the monotonic graphs of fig. 21. It is a well-known theorem that a set of curves contained in a bounded region, and whose lengths are uniformly bounded, is compact.[44a] These conditions are evidently fulfilled by the monotonic graphs, where the uniform bound for the length is the sum of two non-parallel sides of the rectangle.

By composition (diagonal process)[44b] of $g_1, g_2, \cdots, g_k$ and R, it follows that *the set of all proper and improper representations* $[(g, R)]$ *of a given system of contours* (Γ) *is compact.*

---

[44] See (16.4).

[44a] Fréchet [26]. Hilbert, Jahresbericht der Deutschen Mathematiker-Vereinigung, v. 8 (1899), p. 184.

[44b] Cf. Courant-Hilbert, *Methoden der Mathematischen Physik*, 1924, p. 42.

The limit of a variable representation (g, R) of (Γ) is evidently itself a representation (g_0, R_0), proper or improper, of the same contours. Therefore, *the set* $[(g, R)]$ *of all proper and improper representations of a given system of contours* (Γ) *is closed.* This refers to the general definition: an L-set is closed if it contains the limit of every sequence of its elements.

The compactness property extends to the case where (Γ_m, g_m, R_m) is a sequence of representations of a *variable* contour-system $(\Gamma)_m$, provided that $(\Gamma)_m$ tends in the Fréchet sense to a fixed set of contours (Γ_0). For we have referred also this case to the monotonic graphs of fig. 21. Hence a sub-sequence $(\Gamma_{m'}, g_{m'}, R_{m'})$ exists which converges to a representation (Γ_0, g_0, R_0) of the contours (Γ_0).

It is evident that in all these modes of approach, the value of $g_m(z_m)$ converges to that of $g_0(z_0)$, except at most at a denumerable infinity of points—those of discontinuity of g_0. Also $g_m(z_m)$ evidently remains uniformly bounded; therefore we may pass to the limit under a sign of integration provided that any other factors under the integral sign also remain uniformly bounded. Often, the subsistence of the preceding two conditions in the approach of (Γ_m, g_m, R_m) to (Γ_0, g_0, R_0) will be sufficient for our purposes.

4. $A(g, R)$ for improper representations

We must have the definition of $A(g, R)$ on the total set $[(g, R)]$, proper or improper.

The case where R separates has already been considered, (5.22); we simply form the functional A for each component part and take the sum.

Turning to improper parametric representations g, we see that the continuously improper case (a) requires no discussion.

For improper representations of the discontinuous type (b), we easily show that

$$A(g, R) = +\infty. \tag{13.1}$$

In fact, let the transition curve C_j be represented, e.g., as a circle, and let the discontinuity occur at the point P of this circle. Let Q_1, Q_2 denote two points on C_j near to P and on opposite sides of it. Suppose Q_1, Q_2 joined by a circular arc $Q_1Q_3Q_2$ lying in the interior of the circular region which represents the semi Riemann surface R'. Then $Q_1Q_3Q_2$ and the arc Q_1PQ_2 bound a simply-connected region ρ, which by an elementary conformal transformation can be converted into a circle γ. The Dirichlet integral over γ of the transformed harmonic function H is known to be $+\infty$, on account of the presumed type of discontinuity in the boundary values, i.e. with distinct one-sided limits.[45] Therefore $D(H) = +\infty$ over ρ, since the Dirichlet integral is invariant under conformal transformation. Therefore $D(H)$ itself, $\geqq D(H)$ over ρ, is equal to $+\infty$.

[45] See [5], pp. 298–300.

In short, the situation here is a local one, referring only to the immediate neighborhood of the point of discontinuity P, and can therefore be reduced to the corresponding situation for the one-contour case.

Quite similar reasoning applies to the type of improper representation where the algebraic curve corresponding to R has a real double point. As has been stated, this may be represented also by two circular boundaries C_1, C_2 tangent at a point P. Since, by hypothesis, the contours Γ_1, Γ_2 have no common point, the vector g has two distinct limits at P according as we approach along C_1 or C_2. Therefore, $D(H)$ taken over the part of R' that lies within a distance ϵ of P is $+\infty$; consequently, over all of R', we have $D(H) = A(g, R) = +\infty$. A real multiple point of any order can be treated similarly.

More delicate is the question of how to define $A(g, R)$ for the degenerate representations. For instance, let the point α of Γ_1 correspond to all of C_1, while the representations $g_2, \cdots, g_k$ of the other contours are non-degenerate. Let H denote the harmonic vector on R' with the boundary values $\alpha, g_2, \cdots, g_k$. Then it will not do to define, literally, $A(g, R) = D(H)$. Rather, the following definition:

$$A(g, R) = D(H) + d(\Gamma_1), \tag{13.2}$$

is the one that will be made, being definitely indicated by the lower semi-continuity property of $A(g, R)$, next to be considered. Here, as defined in §5, $d(\Gamma_1) = \min A(g_1)$ refers to simply-connected surfaces bounded by Γ_1.

Similarly, if the representations of several contours degenerate, e.g. Γ_1, Γ_2, Γ_3 into the points α_1, α_2, α_3, we first define H with the boundary valves $\alpha_1, \alpha_2, \alpha_3, g_4, \cdots, g_k$, and then

$$A(g, R) = D(H) + d(\Gamma_1) + d(\Gamma_2) + d(\Gamma_3). \tag{13.3}$$

§14. Lower Semi-Continuity of $A(g, R)$

1. Proof

The functional $A(g, R)$ is *lower semi-continuous*, and this property subsists for all representations proper or improper.

The definition of lower semi-continuous is: if $(g, R) \to (g_0, R_0)$, then

$$A(g_0, R_0) \leqq \underline{\lim}\, A(g, R), \tag{14.1}$$

where lim denotes the inferior limit.

Let $\bar{R}_0'$ be a closed interior sub-region of the semi Riemann surface R_0'; then as soon as the variable R' comes sufficiently near to R_0', $\bar{R}_0'$ coincides also with a closed interior sub-region of R'. By a known theorem, since g remaining bounded tends to g_0, the partial derivatives $\partial H/\partial u$, $\partial H/\partial v$ of the harmonic function H in R' with the boundary values g tend to $\partial H_0/\partial u$, $\partial H_0/\partial v$ uniformly in $\bar{R}_0'$, where H_0 denotes the harmonic function on R_0' with the boundary values g_0.

It follows that

$$D(H_0) \text{ over } \bar{R}_0' = \lim \{D(H) \text{ over } \bar{R}_0'\}.$$

But

$$D(H) \text{ over } \bar{R}_0' \leqq D(H).$$

Therefore

$$D(H_0) \text{ over } \bar{R}_0' \leqq \underline{\lim}\, D(H).$$

Allow $\bar{R}_0'$ to approach to R_0' ; then

$$D(H_0) \text{ over } \bar{R}_0' \to D(H_0);$$

therefore

$$D(H_0) \leqq \underline{\lim}\, D(H), \tag{14.2}$$

which is the stated lower semi-continuity property.

2. The inequality

$$d(\Gamma, r) \leqq d(\Gamma, r'). \tag{14.3}$$

Here the first member refers to proper representations, the second to those of the primary reduced type. We shall prove also

$$d(\Gamma, r') = d(\Gamma, r''), \tag{14.4}$$

which justifies the use of only the primary r' instead of the general reduced type r''.

A primary reduction of R may be, first, a separation into two components: $R = R_1 + R_2$, with a corresponding partition of the characteristic: $r = r_1 + r_2$, of the transition curves: $(C) = (C)_1 + (C)_2$, and of the related contours: $(\Gamma) = (\Gamma)_1 + (\Gamma)_2$.

We can form R_1, R_2, and then build up to R in the following manner. Let R_1, R_2 be chosen, and also parametric representations g_1, g_2 of the contours $(\Gamma)_1$, $(\Gamma)_2$ respectively, so that

$$A(g_1, R_1) < d(\Gamma_1, r_1) + \epsilon, \tag{14.5}$$

$$A(g_2, R_2) < d(\Gamma_2, r_2) + \epsilon. \tag{14.6}$$

Of course, this is possible by definition of $d(\Gamma_1, r_1)$, $d(\Gamma_2, r_2)$.

Now, by means of a branch-cut between two arbitrarily near points α, $\alpha + \delta$, both belonging to a sheet of the semi-surface R_1' and a sheet of the semi-surface R_2', we can unite R_1' and R_2' to form the semi Riemann surface R', with $r = r_1 + r_2$.[45a] The boundaries of R' are those of R_1', R_2' combined, and we keep on these the representations g_1, g_2, so forming a representation g of (Γ).

Then in the formula (9.11) for $A(g, R)$, the parameter δ enters only in the

[45a] We slit the sheet of R_1' and that of R_2' between the points $z = \alpha$, $z = \alpha + \delta$, and then join the opposite edges of the slits crosswise in the usual way.

"kernel" $P(z, \zeta)$, and in an analytic way; consequently $A(g, R)$ is an analytic function[46] of δ, therefore certainly a *continuous* one. We have for $\delta = 0$,

$$A(g, R) = A(g_1, R_1) + A(g_2, R_2);$$

therefore for δ near enough to zero,

$$A(g, R) < A(g_1, R_1) + A(g_2, R_2) + \epsilon < d(\Gamma_1, r_1) + d(\Gamma_2, r_2) + 3\epsilon,$$

by (14.5, 6). Accordingly, since R' is properly of the topological type r,

$$d(\Gamma, r) \leqq A(g, R) < d(\Gamma_1, r_1) + d(\Gamma_2, r_2) + 3\epsilon.$$

Allowing ϵ to approach zero, we get

$$d(\Gamma, r) \leqq d(\Gamma_1, r_1) + d(\Gamma_2, r_2), \tag{14.7}$$

which is (14.3) for the type of reduction of R considered.

The other type of primary reduction of R consists of a reduction of the characteristic r of its semi-surface R' to $r - 2$ or $r - 1$, according as R' is two-sided or one-sided. Keeping R symmetric, we can increase the characteristic of R' by *two* by means of an additional branch-cut from α to $\alpha + \delta$, together with the conjugate branch-cut $\bar{\alpha}$, $\bar{\alpha} + \bar{\delta}$. In the one-sided case, we can adjoin a branch-cut $a - i\delta$, $a + i\delta$, where a and δ are *real*, to increase the characteristic by *one*, and render the surface one-sided, if it is not so already. Conversely, if we allow $\delta \to 0$, R undergoes a primary reduction of its characteristic. Again $A(g, R)$ is a continuous function of δ for $\delta \to 0$, and the same reasoning applies.

We have, in sum, the desired relation (14.3) for every type of primary reduction.

The second relation, (14.4), is proved in the same way. For by a similar construction, we can build up any composite reduced Riemann surface into a primary reduced one, arbitrarily near to it in form. Reversing the process, the primary reduced surface is expressed as limit of a composite reduced surface. It follows, as before, that

$$d(\Gamma, r') \leqq d(\Gamma, r'').$$

But, according to our phrasing of the definition, the primary type r' is part of the general type r''. Therefore, we also have the reverse inequality, and hence the equality (14.4).

For the same reason, we can also express the relations (14.3, 4) in the form

$$d(\Gamma, r) = d(\Gamma, r, r''); \tag{14.8}$$

i.e. the lower bound of $A(g, R)$ for the proper type r of Riemann surface R is also the lower bound for all types, proper and reduced.

[46] Generally, this function involves a term in $1/\log \delta$ (cf. [3], p. 339, formulas 5.5, 5.6), but evidently this does no harm to our continuity argument.

3. Lower semi-continuity of $d(\Gamma, r)$[47]

We have defined

$$d(\Gamma, r) = \min A(g, R), \tag{14.9}$$

with respect to all parametric representations g of the given set of contours (Γ), and all *proper* symmetric Riemann surfaces of characteristic r. It is easy to prove the lower semi-continuity of this functional.

Let the sequence $(\Gamma)_m$ tend to (Γ) in the Fréchet sense. Then $(\Gamma)_m$ must remain in a bounded region of space.

By definition of lower bound, we can find for every $\epsilon_m > 0$ a representation (g_m, R_m) of $(\Gamma)_m$ such that

$$d(\Gamma_m, r) \leqq A(g_m, R_m) < d(\Gamma_m, r) + \epsilon_m . \tag{14.10}$$

Let us arrange that $\epsilon_m \to 0$ as $m \to \infty$; then

$$\underline{\lim}\, d(\Gamma_m, r) \leqq \underline{\lim}\, A(g_m, R_m) \leqq \underline{\lim}\, d(\Gamma_m, r),$$

that is

$$\underline{\lim}\, d(\Gamma_m, r) = \underline{\lim}\, A(g_m, R_m). \tag{14.11}$$

By definition of $\underline{\lim}$, we can select a sub-sequence m', such that

$$\lim A(g_{m'}, R_{m'}) = \underline{\lim}\, A(g_m, R_m). \tag{14.12}$$

By the compactness of (g_m, R_m), which applies, according to §13.3, because $(\Gamma)_m \to (\Gamma)$, we can in turn select a sub-sequence m'' such that

$$(g_{m''}, R_{m''}) \to (g, R),$$

where (g, R) denotes some proper or improper representation of (Γ). Whichever is the case, we have by (14.8),

$$d(\Gamma, r) \leqq A(g, R). \tag{14.13}$$

By (14.12, 11),

$$\lim A(g_{m''}, R_{m''}) = \lim A(g_{m'}, R_{m'}) = \underline{\lim}\, A(g_m, R_m) = \underline{\lim}\, d(\Gamma_m, r). \tag{14.14}$$

But, by the lower semi-continuity of $A(g, R)$,

$$A(g, R) \leqq \lim A(g_{m''}, R_{m''}). \tag{14.15}$$

Therefore, combining the last three formulas, we have

$$d(\Gamma, r) \leqq \underline{\lim}\, d(\Gamma_m, r), \tag{14.16}$$

which was to be proved.

---

[47] This fact is essentially contained in the discussion of [5], pp. 319–320.

4. Degenerate representations and lower semi-continuity

The apparently arbitrary definition (13.2) is uniquely determined by the lower semi-continuity of $A(g, R)$. That is, if (g_0, R_0) is a degenerate representation, and (g, R) a variable proper representation tending to (g_0, R_0), then with the definition (13.2)of $A(g_0, R_0)$, we have

$$A(g_0, R_0) = \min \lim A(g, R), \tag{14.17}$$

the minimum being with respect to all modes of approach of (g, R) to (g_0, R_0).

To fix ideas, let there be three contours $\Gamma_1, \Gamma_2, \Gamma_3$ represented topologically on three circles C_1, C_2, C_3. Let the representation of Γ, on C_1 be degenerate in that all of C_1 corresponds to the point α of Γ_1, while all of Γ_1 corresponds to the point P of C_1. We shall suppose that the other two parametric representations g_2^0, g_3^0 are non-degenerate.

Let R_0 denote the circular region $C_1C_2C_3$, and for simplicity, let R first remain constantly the same as R_0, while $g = (g_1, g_2, g_3)$ tends to $g_0 = (\alpha, g_2^0, g_3^0)$. Let

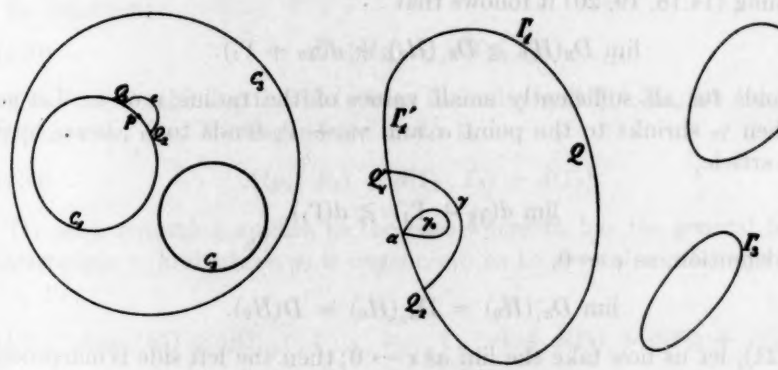


FIG. 23

H denote the harmonic vector in R with the boundary values g, and H_0 the harmonic vector in $R_0(=R)$ with the boundary values g_0.

Suppose that about P as center a small circular arc c or Q_1cQ_2 is constructed. Let H convert the arc Q_1PQ_2 of C_1 into the arc Γ_1' or $Q_1'P'Q_2'$ of Γ_1, and the circular arc Q_1cQ_2 into the arc γ joining $Q_1'Q_2'$. By hypothesis, when $g \to g_0$, Γ_1' tends to cover all of Γ_1, while γ tends to an arc γ_0 with both end-points at the point α. Thus the closed curve $\gamma + \Gamma_1'$ tends to the closed curve $\gamma_0 + \Gamma_1$.

Let c divide R into two regions R_1 and ρ, the latter being the small region close to P. Then we have

$$D_R(H) = D_{R_1}(H) + D_\rho(H).$$

Therefore, as $g \to g_0$,

$$\underline{\lim}\, D_R(H) \geqq \underline{\lim}\, D_{R_1}(H) + \underline{\lim}\, D_\rho(H). \tag{14.18}$$

By the lower semi-continuity of the Dirichlet integral,

$$\underline{\lim}\, D_{R_1}(H) \geqq D_{R_1}(H_0). \tag{14.19}$$

By definition,

$$d(\gamma + \Gamma_1') \leqq D_\rho(H).$$

Therefore

$$\underline{\lim}\, d(\gamma + \Gamma_1') \leqq \underline{\lim}\, D_\rho(H).$$

But since $\gamma + \Gamma_1' \to \gamma_0 + \Gamma_1$, we have by the previous article

$$d(\gamma_0 + \Gamma_1) \leqq \underline{\lim}\, d(\gamma + \Gamma_1'),$$

hence

$$d(\gamma_0 + \Gamma_1) \leqq \underline{\lim}\, D_\rho(H). \tag{14.20}$$

Combining (14.18, 19, 20) it follows that

$$\underline{\lim}\, D_R(H) \geqq D_{R_1}(H_0) + d(\gamma_0 + \Gamma_1). \tag{14.21}$$

This holds for all sufficiently small values of the radius ϵ of c. Let now $\epsilon \to 0$; then γ_0 shrinks to the point α, and $\gamma_0 + \Gamma_1$ tends to Γ_1; hence by the previous article,

$$\underline{\lim}\, d(\gamma_0 + \Gamma_1) \geqq d(\Gamma_1). \tag{14.22}$$

Also, by definition, as $\epsilon \to 0$,

$$\lim D_{R_1}(H_0) = D_{R_0}(H_0) = D(H_0). \tag{14.23}$$

In (14.21), let us now take the $\underline{\lim}$ as $\epsilon \to 0$; then the left side is independent of ϵ, and by the last two formulas we get

$$\underline{\lim}\, D_R(H) \geqq D(H_0) + d(\Gamma_1), \tag{14.24}$$

where the $\underline{\lim}$ is for $g \to g_0$.

Evidently, similar reasoning applies if the circular region R is variable and tends to R_0.

Accordingly, if we make the definition (13.2), i.e.

$$A(g_0, R_0) = D(H_0) + d(\Gamma_1), \tag{14.25}$$

then (14.24) states that

$$A(g_0, R_0) \leqq \underline{\lim}\, A(g, R); \tag{14.26}$$

i.e. the lower semi-continuity of $A(g, R)$ is preserved by this definition for the degenerate representations.

That $A(g_0, R_0)$ is exactly the minimum limit, in accordance with (14.17), can be seen by keeping R_0 and g_2^0, g_3^0 fixed, while allowing g_1 to be a *minimizing* representation of Γ_1: $A(g_1) = d(\Gamma_1)$, that tends to degeneration. This is

possible because of the three parameters involved in the minimizing representation of $A(g)$.[48]

The extension to degenerate representations of several contours, e.g. (13.3), is obvious.

COROLLARY. We deduce from (14.25) the following corollary, useful for the sequel. Let H_0 be continued as the constant α into the interior of C_1. Then H_0 and α form a piece-wise continuously differentiable function φ in the ring R_{23} between C_2 and C_3, and we have

$$D_{23}(\varphi) = D_0(H_0), \tag{14.27}$$

where the region of integration is R with the indicated subscripts. Let H_{23} denote the harmonic function in the ring R_{23} with the same boundary values g_1^0, g_3^0 as φ; then

$$D_{23}(H_{23}) \leqq D_{23}(\varphi), \tag{14.28}$$

by the minimum property of harmonic functions in the Dirichlet integral.

By definition,

$$D_{23}(H_{23}) \geqq d(\Gamma_2, \Gamma_3). \tag{14.29}$$

Therefore, combining the last three formulas and (14.25), we have for any representation (g_0, R_0) degenerate as to Γ_1:

$$A(g_0, R_0) \geqq d(\Gamma_2, \Gamma_3) + d(\Gamma_1). \tag{14.30}$$

The same reasoning applies to the case where R_0 has the general topological characteristic r, and where g_0 is degenerate as to several controus, e.g. $\Gamma_1, \Gamma_2, \cdots, \Gamma_j$:

$$A(g_0, R_0) \geqq d(\Gamma_{j+1}, \Gamma_{j+2}, \cdots, \Gamma_k; r) + d(\Gamma_1) + d(\Gamma_2) + \cdots + d(\Gamma_j). \tag{14.31}$$

Since the right member represents a particular mode of reduction, we have therefore for any degenerate representation:[49]

$$A(g_0, R_0) \geqq d(\Gamma, r'). \tag{14.32}$$

§15. COMPLETION OF THE SOLUTION

1. Attainment of the minimum of $A(g, R)$

In summary, the set of *all* representations $[(g, R)]$, *proper and improper*, of a given system (Γ, r) is *compact* and *closed*, and $A(g, R)$ is a *lower semi-continuous* functional on this set.

Now, we have the following general theorem:[50]

Any lower semi-continuous functional on a compact and closed set attains its minimum value.

---

[48] See [5], p. 277.

[49] If R_0 is proper, we can even assert the strict relation $>$ in (14.32). For φ in (14.27) is certainly non-harmonic, if C_1 does not reduce to a point.

[50] Fréchet [26].

The proof is simply as follows. Let $U(a)$ denote a lower semi-continuous functional on the compact and closed set $[a]$. If d denote the lower bound of $U(a)$, it is always possible to construct a "minimizing sequence" $\{a_m\}$, i.e. one such that

$$\lim U(a_m) = d.$$

This construction is merely an application of the definition of lower bound. Since $[a]$ is compact, $\{a_m\}$ contains a convergent sub-sequence $\{a_{m'}\}$; and since $[a]$ is closed, the limit of $\{a_{m'}\}$ is an element a^* of $[a]$:

$$\lim a_{m'} = a^*.$$

By the postulated lower semi-continuity,

$$U(a^*) \leqq \underline{\lim}\, U(a_{m'}) = \lim U(a_m) = d.$$

But of the alternatives $U(a^*) \leqq d$, the former is contrary to the definition of lower bound; therefore $U(a^*) = d$, which was to be proved.

The immediate application is that $A(g, R)$ *attains its minimum*:

$$A(g^*, R^*) = d(\Gamma, r) \tag{15.1}$$

for some representation (g^*, R^*).

2. Exclusion of improper representations

However, to secure the property of closure of the set $[(g, R)]$ essential to the preceding reasoning, we have been compelled to adjoin various types of improper representations. We must now show that (g^*, R^*) is not one of these, but rather a proper representation.

We make, according to the statement of Theorem II, the two hypotheses

$$d(\Gamma, r) < +\infty, \tag{15.2}$$

$$d(\Gamma, r) < d(\Gamma, r'). \tag{15.3}$$

First, we show that R^* cannot be of the reduced type. For then

$$A(g^*, R^*) \geqq d(\Gamma, r'') = d(\Gamma, r') > d(\Gamma, r), \tag{15.4}$$

contrary to (15.1).

For the same reason, we see, by reference to (14.32), that g^* cannot be degenerate.

Neither can g^* be discontinuously improper. For then, by (13.1), we would have

$$A(g^*, R^*) = +\infty, \tag{15.5}$$

contrary to (15.2) and (15.1).

The same reason shows that R^* cannot be improper in virtue of two transition curves having a point in common, i.e. the corresponding algebraic curve cannot

have a real multiple point. For, as we have seen in connection with (13.1), this also leads to the impossible condition (15.5).

Therefore R^* is a proper Riemann surface of type (Γ, r), and g^* is proper, or at worst continuously improper.

3. The required minimal surface exists

Let

$$x_i = H_i^*(u, v) = \Re F_i(w) \qquad (i = 1, 2, \cdots, n) \tag{15.6}$$

denote the harmonic functions on the semi-surface $R^{*\prime}$ of R^* with the continuous values $g_i^*(z)$ on the boundary curves C. H^* or F is determined explicitly by the formula (8.1) or (8.21).

Then, on account of the minimum value of $A(g^*, R^*)$, we have the variational condition

$$\delta A(g, R) = 0 \tag{15.7}$$

for $(g, R) = (g^*, R^*)$. As has been shown in formula (12.29), this implies

$$\sum_{i=1}^{n} F_i'^2(w) = 0. \tag{15.8}$$

Consequently, the surface M defined by the equations (15.6) is a *minimal surface of the prescribed topological type bounded by the given set of contours* (Γ).

Thus Theorem II is proved, and the Platean problem is solved in the stated generality.

4. The correspondence between boundaries is one-one

It remains only to exclude the continuously improper representations—those in which a partial arc of a transition curve C may correspond to a single point of the related contour Γ.

If g^* is constant on a segment σ of a real branch-cut of R^*, the harmonic vector function H^* can be prolonged by Schwarzian symmetry over σ:

$$H^*(u, -v) = -H^*(u, v). \tag{15.9}$$

Since

$$F_i'(w) = \frac{\partial H_i^*}{\partial u} - i\frac{\partial H_i^*}{\partial v}, \tag{15.10}$$

the condition (15.8) gives, when we take its real part,

$$\sum_{i=1}^{n}\left(\frac{\partial H_i^*}{\partial u}\right)^2 = \sum_{i=1}^{n}\left(\frac{\partial H_i^*}{\partial v}\right)^2. \tag{15.11}$$

Because $H_i^*(u, v)$ is constant on σ, we have on σ:

$$\frac{\partial H_i^*}{\partial u} = 0, \qquad i = 1, 2, \cdots, n. \tag{15.12}$$

From this and (15.11), it follows that on σ:

$$\frac{\partial H_i^*}{\partial v} = 0, \qquad i = 1, 2, \cdots, n. \tag{15.13}$$

Consequently, on σ,

$$F_i'(w) = 0, \qquad i = 1, 2, \cdots, n. \tag{15.14}$$

But, by the Schwarzian symmetry, σ is *interior* to a domain of regularity of each $F_i(w)$. Therefore

$$F_i'(w) \equiv 0, \qquad i = 1, 2, \cdots, n; \tag{15.15}$$

hence

$$H_i^*(u, v) = \Re F_i(w) \equiv c_i, \quad \text{a constant}, \tag{15.16}$$

for $i = 1, 2, \cdots, n$.

But this is absurd, since, by hypothesis, H^* is not even constant on the entire transition curve C, but only on a *partial* segment σ. Moreover, (15.16) would imply that the set of contours (Γ) reduces to a point, which is absurd.

§16. Least Area

1. The Lebesgue definition of area

With all the advantages that result by the introduction of the vector Dirichlet functional $D(\mathfrak{x})$ or the equivalent functional $A(g, R)$, into the Plateau problem, we must not lose sight of the fact that, after all, the interest of the problem lies in the least area property of its solution M as compared with all surfaces of the same topological type with the same boundaries.

To establish this least area property, we adopt as definition of area the one given (for simply-connected surfaces) by Lebesgue:[51]

$$\mathfrak{A}(S) = \min \varliminf_{m \to \infty} \mathfrak{S}(\Pi_m). \tag{16.1}$$

Here S is any given surface, Π_m any sequence of polyhedral surfaces of the same topological type as S, which approach to S as $m \to \infty$. $\mathfrak{S}(\Pi_m)$ denotes the area of Π_m in the elementary sense, and the minimum is with respect to all sequences Π_m.

The approach of Π_m to S will be understood in the Fréchet sense:[52]

$$\text{dist}\,(\Pi_m S) \to 0 \quad \text{as} \quad m \to \infty. \tag{16.2}$$

Let T denote any topological correspondence[53] between Π_m and S, in which P, Q are any corresponding points; then the Fréchet definition is

$$\text{dist}\,(\Pi_m S) = \min_T \max_{(P,Q)} \overline{PQ}, \tag{16.3}$$

---

[51] See footnote 27.

[52] [26].

[53] This is defined in all cases by establishing a topological correspondence, respecting symmetric point-pairs, between the Riemann surfaces on which Π_m and S are represented.

where $\overline{PQ}$ denotes the ordinary distance between two points, and, as indicated, the maximum is with respect to all point-pairs P, Q which correspond by T, while the minimum is with respect to all topological correspondences T.

We apply a similar definition to the boundaries $(\Gamma)_m$, (Γ) of Π_m, S respectively, and by $(\Gamma)_m \to (\Gamma)$ we understand, following Fréchet,

$$\text{dist}\,(\Gamma_m\Gamma) \to 0 \quad \text{as} \quad m \to \infty. \tag{16.4}$$

This condition is obviously a consequence of (16.2), since in any topological correspondence T between Π_m and S, the respective boundaries must also correspond topologically, so that

$$\text{dist}\,(\Gamma_m\Gamma) \leqq \text{dist}\,(\Pi_m S). \tag{16.5}$$

2

In proving the existence of the minimal surface M, we made it a point not to use anything of the theory of conformal mapping, basing our treatment rather on the functional $A(g, R)$. In fact, in §18, we shall even use this functional as a means of establishing some of the classic theorems concerning the conformal mapping of plane regions.

But to deal with the least area property of M, we now employ the theorem[54] that it is possible to map Π_m conformally on a semi Riemann surface R_m' of the same topological type as Π_m. The map remains one-one and continuous, even piece-wise analytic, on the boundaries.[55]

We have seen[56] that in a conformal representation of any surface, the vector Dirichlet integral is equal to the area integral. Therefore, denoting also by Π_m the vector function on R_m' which defines the conformal map, we have

$$\mathfrak{S}(\Pi_m) = D(\Pi_m). \tag{16.6}$$

As stated, the function Π_m determines on the boundaries C_m of R_m' a parametric representation g_m of its own boundaries Γ_m. If H_m denote the harmonic vector on R_m' with the boundary values g_m, we have, by the minimum property of the Dirichlet integral:

$$A(g_m, R_m) = D(H_m) \leqq D(\Pi_m). \tag{16.7}$$

Consequently,

$$d(\Gamma_m, r) \leqq \mathfrak{S}(\Pi_m), \tag{16.8}$$

where r denotes the prescribed topological characteristic of Π_m.

In other words: for any given set of *polygonal* contours and prescribed topological type r, the lower bound of the Dirichlet integral is $\leqq$ the area of any *polyhedral* surface of type r bounded by these contours.

---

[54] Due to Schwarz, for $k = 1$, $r = 0$; for the higher cases, part of the general theory of conformal mapping of reimannian manifolds, as developed largely by P. Koebe.

[55] According to the Schwarz reflection principle.

[56] Theorem (a), (5.24).

3. The area of M, as an integral, and according to Lebesgue

We next observe that if we consider the area of M *as defined by an integral*:

$$I(M) = \int\int \sqrt{EG - F^2}\, du\, dv, \tag{16.9}$$

we have

$$I(M) = D(H^*) = A(g^*, R^*) = d(\Gamma, r). \tag{16.10}$$

By Theorem (a), §5, this follows simply from the condition for a minimal surface:

$$\sum_{i=1}^{n} F_i'^2(w) = (E - G) - 2iF = 0,$$

which implies that the representation of M on R^* is conformal:

$$E = G, \qquad F = 0.$$

For a surface as regular as M (analytic), it is known that the area $I(M)$, as defined by the integral (16.9), is the same as the Lebesgue area.[56a] However, in order that our presentation may be as self-contained as possible, we shall not use this fact, but only the weaker and much more easily proved relation

$$\mathfrak{A}(M) \leqq I(M). \tag{16.11}$$

The one-contour case will provide a sufficiently typical illustration of the proof. Let C denote the unit circle and C_ρ a concentric circle of radius $\rho < 1$. Then by taking ρ near enough to unity, and constructing a sufficiently fine polar coördinate net over C_ρ, we obtain a polyhedral surface Π inscribed in M, which is arbitrarily near to M in the Fréchet sense, and whose area is arbitrarily near to $I(M)$. (Each mesh of the polar net gives, when subdivided by a diagonal, two triangular faces of Π.) The proof proceeds by showing that as the fineness of the net is increased, ρ being fixed, we have $\mathfrak{S}(\Pi) \to I_\rho(M)$, where the latter denotes the area integral over C_ρ. This is easily accomplished along the lines of the corresponding proof for the length integral,[56b] and is valid under the same hypothesis: that M be continuously differentiable.

Thus we can construct a sequence of polyhedra Π tending to M, and whose areas tend to $I(M)$. By the definition (16.1) of the Lebesgue area as a *minimum* limit, the result (16.11) follows.

4. The area of M is a minimum

Let now S denote any continuous surface of topological type r bounded by (Γ). Then, as is possible by definition (16.1), let a sequence of polyhedral surfaces Π_m be constructed, which tend to S and are such that

$$\underline{\lim}\, \mathfrak{S}(\Pi_m) = \mathfrak{A}(S). \tag{16.12}$$

[56a] See footnote 28.

[56b] See [28], pp. 1–2.

Using the lower semi-continuity of $d(\Gamma, r)$ and the result (16.8), we infer

$$d(\Gamma, r) \leqq \underline{\lim}\, d(\Gamma_m, r) \leqq \underline{\lim}\, \mathfrak{E}(\Pi_m) = \mathfrak{A}(S). \tag{16.13}$$

S may be taken to be M; therefore, with the use of (16.11, 10), 23 have

$$d(\Gamma, r) \leqq \mathfrak{A}(M) \leqq I(M) = d(\Gamma, r). \tag{16.14}$$

Consequently,

$$d(\Gamma, r) = \mathfrak{A}(M) = I(M). \tag{16.15}$$

It follows by (16.13) that

$$\mathfrak{A}(M) \leqq \mathfrak{A}(S), \tag{16.16}$$

which is the desired minimum property of the area of M.

Incidentally, in (16.15) the two interpretations of the area of M are shown to be equivalent.

According to (16.16), we have with the definition (5.1)

$$a(\Gamma, r) = \min \mathfrak{A}(S) = \mathfrak{A}(M); \tag{16.17}$$

therefore by (16.15),

$$a(\Gamma, r) = d(\Gamma, r). \tag{16.18}$$

This is the equivalence of the minima of area and of the Dirichlet functional stated in (5.26).

The results of this section are embodied in Theorem I (§5).

§17. Arbitrary Jordan Contours

1

In Theorems I and II, the contours (Γ) are subject to the restriction

$$a(\Gamma, r) = d(\Gamma, r) < + \infty. \tag{17.1}$$

We now lift this restriction, and allow the contours Γ to be perfectly *general* Jordan curves, where usually[57]

$$a(\Gamma, r) = d(\Gamma, r) = + \infty. \tag{17.2}$$

The sufficient condition of Theorems I and II is now replaced by

$$\bar{e}(\Gamma, r) > 0, \tag{17.3}$$

as stated (§5) in Theorem III, which we proceed to prove.

Let $(\Gamma)_m$ denote a sequence of contour systems tending to (Γ), such that $d(\Gamma_m, r)$ is finite for all m, and

$$\lim_{m \to \infty} e(\Gamma_m, r) = \lim_{m \to \infty} \{d(\Gamma_m, r') - d(\Gamma_m, r)\} = \bar{e}(\Gamma, r). \tag{17.4}$$

[57] Cf. §5.5.

The existence of $(\Gamma)_m$ is implied by the definition of $\bar{e}(\Gamma, r)$.

By (17.3), it follows that at least from a certain value of m on,

$$d(\Gamma_m, r) < d(\Gamma_m, r'). \tag{17.5}$$

Hence, omitting only a finite number of values of m at the start, we have by Theorem II that there exists for every m a representation (g_m^*, R_m^*) of the contours $(\Gamma)_m$, where R_m^* is of topological type r, such that: if

$$x_i = \Re F_{mi}(w)$$

are the harmonic functions on the semi-surface of R^* with the boundary values $g_{mi}^*(z)$, then

$$\sum_{i=1}^{n} F_{mi}'^2(w) = 0. \tag{17.6}$$

The set $[(g_m^*, R_n^*)]$ is compact, since the contours $(\Gamma)_m$ are approaching to the contours (Γ). Therefore we can select a partial sequence—denote it without change of the index m—such that (g_m^*, R_m^*) tends to a representation (g^*, R^*), proper or improper, of the contours (Γ). The semi-surface of R^* is then of topological type r or one of the reduced types r''.

2. R^* is proper

We prove first that $R^{*\prime}$ is properly of type r.

For suppose the contrary. Then as $m \to \infty$, R_m^* tends to separate, or its characteristic to reduce, by the disappearance of certain branch-cuts β_m. When m is sufficiently large, these branch-cuts are therefore arbitrarily short. Let us then modify R_m^* while m is still finite, by severing the branch-cuts β_m and then re-uniting the opposite edges in their own sheets, instead of cross-wise as is done to form R_m^*. This produces a Riemann surface $\tilde{R}_m^*$, which is either separated or reduced in its characteristic exactly like the limit surface R^*.

On the transition curves of $\tilde{R}_m^*$, which evidently have not been disturbed in the transformation from R_m^*, we keep the same parametric representation g_m^* of the contours $(\Gamma)_m$. Then we form[58]

$$B(g_m^*, R_m^*, \tilde{R}_m^*) = A(g_m^*, \tilde{R}_m^*) - A(g_m^*, R_m^*). \tag{17.7}$$

Since, evidently,

$$A(g_w^*, \tilde{R}_m^*) \geqq d(\Gamma_m, r'), \qquad A(g_m^*, R_m^*) = d(\Gamma_m, r), \tag{17.8}$$

we have

$$B(g_m^*, R_m^*, \tilde{R}_m^*) \geqq d(\Gamma_m, r') - d(\Gamma_m, r) = e(\Gamma_m, r) \geqq 0. \tag{17.9}$$

Next we show that if R_m^* tends to separate as assumed, we have

$$\lim_{m \to \infty} B(g_m^*, R_m^*, \tilde{R}_m^*) = 0. \tag{17.10}$$

[58] Cf. the two-contour case [3], pp. 354–356.

We can transform the definition (17.7) by the formula (9.11), getting

$$B(g_m^*, R_m^*, \tilde{R}_m^*) = \frac{1}{4\pi}\int_C\int_C [g_m^*(z) - g_m^*(\zeta)]^2\{\tilde{P}_m(z,\zeta) - P_m(z,\zeta)\}\,dzd\zeta. \tag{17.11}$$

Here $P_m(z,\zeta)$ is $P(z,\zeta)$ for the Riemann surface R_m^*, and $\tilde{P}_m(z,\zeta)$ is a function on $\tilde{R}_m^*$ defined as follows: if $\tilde{R}_m^*$ does not separate, $\tilde{P}_m(z,\zeta)$ is simply $P(z,\zeta)$ for that surface; if $\tilde{R}_m^*$ separates, and z, ζ lie on the same component, $\tilde{P}_m(z,\zeta)$ is $P(z,\zeta)$ for that component—if z, ζ lie on different components, $\tilde{P}_m(z,\zeta) = 0$. Since $P_m(z,\zeta)$ and $\tilde{P}_m(z,\zeta)$ both have the same singularity[59] $1/(z-\zeta)^2$ for $z = \zeta$, these singularities neutralize one another in the bracket under the integral sign. Hence the functional B, as expressed by (17.11), is always *well-defined* and *finite-valued*, regardless of whether the functional A has a finite or infinite value. Furthermore, B is a *continuous* functional, since with boundedness of the integrand we may pass to the limit under the integral sign. In the present case,

$$\lim_{m\to\infty} B(g_m^*, R_m^*, \tilde{R}_m^*) = B(g^*, R^*, \tilde{R}^*). \tag{17.12}$$

But, because R^* is supposed to separate or to reduce as assumed, we have that $\tilde{R}^*$ is identical with R^*; therefore according to (17.11),

$$B(g^*, R^*, \tilde{R}^*) = B(g^*, R^*, R^*) = 0. \tag{17.13}$$

Thus (17.10) is proved, and combining it with (17.9) we have

$$\lim_{m\to\infty} e(\Gamma_m, r) = 0; \tag{17.14}$$

consequently by (17.4),

$$\bar{e}(\Gamma, r) = 0, \tag{17.15}$$

which is contrary to the hypothesis (17.3).

Therefore R^* cannot reduce, but is a proper symmetric Riemann surface, which is of type r, as a semi-surface.

3. g^* cannot be improper

With the boundary values $g^*(z)$, let us form on the semi-surface of R^* the harmonic functions

$$x_i = H_i^*(u, v) = \Re F_i(w). \tag{17.16}$$

Then since $g_m^*(z)$ tends to $g^*(z)$ boundedly, we can pass to the limit under the sign in the formula (8.25):

$$F'_{mi}(w) = \frac{1}{\pi i}\int_C g_{mi}^*(z)P_m(z,w)\,dz, \tag{17.17}$$

i.e.

$$\lim_{m\to\infty} F'_{mi}(w) = F'_i(w). \tag{17.18}$$

[59] As remarked in connection with (9.11).

Since (17.6) is obeyed for every value of m, it follows that

$$\sum_{i=1}^{n} F_i'^2(w) = 0. \tag{17.19}$$

Therefore the surface defined by (17.16) is a *minimal surface.*

It remains to be shown that the representation g^* is proper.

g^* cannot be constant on either a partial or complete arc of any transition curve C of R^*. For, by the reasoning of §15, art. 4, it would then follow that H^* is identically a constant; therefore that the same is true of its boundary values g^*. But this is absurd, since the given contours (Γ) certainly do not shrink to a point. Hence both the continuously improper and the degenerate representations are excluded.

[60] Consider next the possibility that g^* be discontinuously improper, by having unequal limits α and β on opposite sides of a point P of a transition curve C, which we may picture as a circle. Then we can construct a small simply-connected region about P, such as ρ in fig. 23. Let ρ be represented conformally on the unit circle,[61] so that P goes over into the point I (1, 0). Since (17.19) is obviously invariant under conformal transformation, we now have in the unit circle again n functions $\tilde{F}_i(w)$ obeying

$$\sum_{i=1}^{n} \tilde{F}_i'^2(w) = 0, \tag{17.20}$$

and where the boundary values of $\Re\tilde{F}_i(w)$ have, on opposite sides of the point $I(1, 0)$, limits α_i, β_i such that

$$l^2 = \sum_{i=1}^{n} (\alpha_i - \beta_i)^2 > 0.^{61a} \tag{17.21}$$

Now take the point P_ϵ: $(1 + \epsilon, 0)$, and let C_ϵ denote the circle with P_ϵ as center and radius equal to the length of the tangent from P_ϵ to the unit circle. Perform the inversion in the circle C_ϵ, followed by reflection in the real axis. The combination is a linear fractional transformation T_ϵ:

$$w' = \frac{aw + b}{cw + d'} \tag{17.22}$$

which evidently transforms the unit circle into itself, and the functions $\tilde{F}_i(w)$ into $\tilde{F}_i(w, \epsilon)$. Again the condition (17.20) is invariant under the analytic transformation of w:

$$\sum_{i=1}^{n} \tilde{F}_i'^2(w, \epsilon) = 0. \tag{17.23}$$

---

[60] Cf. for the following paragraphs [5], pp. 304–306.

[61] Since ρ is bounded by two circular arcs, the requisite conformal transformation can be given explicitly.

[61a] In the condition $l > 0$, the hypothesis is used that each contour Γ is a *Jordan* curve: the end-points α, β of any partial arc of Γ are distinct.

Let $\epsilon \to 0$. Then it is obvious that the tendency of T_ϵ is to spread the boundary values of $\Re \tilde{F}_i(w, \epsilon)$ just above I over the entire lower semi-circumference, and those just below I into the entire upper semi-circumference. These boundary values therefore tend to the constant α_i on the upper semi-circumference and the constant β_i on the lower.

Consequently, as $\epsilon \to 0$, $\Re \tilde{F}_i(w, \epsilon)$ tends to the harmonic function $\Re G_i(w)$ with the boundary values just described. $G_i(w)$ is given explicitly by the equivalent of Poisson's formula:

$$
\begin{aligned}
G_i(w) &= \frac{1}{2\pi} \int_0^{\pi} \alpha_i \frac{e^{i\theta} + w}{e^{i\theta} - w} \, d\theta + \frac{1}{2\pi} \int_{\pi}^{2\pi} \beta_i \frac{e^{i\theta} + w}{e^{i\theta} - w} \, d\theta \\
&= \frac{\alpha_i - \beta_i}{\pi i} \log i \frac{1 + w}{1 + w}.
\end{aligned} \tag{17.24}
$$

Indeed, by Poisson's formula, the complex function $\tilde{F}_i(w, \epsilon)$ itself tends to $G_i(w)$ uniformly in every smaller concentric circle of the unit circle, and therefore we have also for the derivative:

$$
\lim_{\epsilon \to 0} \tilde{F}_i'(w, \epsilon) = G_i'(w). \tag{17.25}
$$

Accordingly, we would have by (17.23),

$$
\sum_{i=1}^{n} G_i'^2(w) = 0. \tag{17.26}
$$

But by actual calculation from (17.24),

$$
\sum_{i=1}^{n} G_i'^2(w) = -\frac{4}{\pi^2} \sum_{i=1}^{n} (\alpha_i - \beta_i)^2 \cdot \frac{1}{(1 - w^2)^2}. \tag{17.27}
$$

According to (17.21), this is in contradiction with (17.26). Thus the proof that $g^*(z)$ cannot be discontinuously improper is completed.

In short, just as in connection with (13.1), the situation here is a local one, referring only to the vicinity of the point of discontinuity P, and can therefore be reduced to the one-contour case.

Finally, by exactly similar reasoning, R^* cannot be improper by the circumstance of two or more of its transition curves having a point in common. Sufficiently typical is the case where two of the bounding circles C_1, C_2 of the circular region representing R^* are tangent at a point P. Then according as we approach P along C_1 or C_2, g^* has two limits α, β, which are *distinct*, $\sum_{i=1}^{n} (\alpha_i - \beta_i)^2 > 0$, since the contours Γ_1, Γ_2 are supposed to have no common points.

In sum, the representation (g^*, R^*) of the contours (Γ) is proper in every respect. The formulas (17.16) therefore represent a minimal surface M of topological type r bounded by (Γ).

4. The minimum area property of M

In general, the area of M is $+\infty$, as well as that of every surface S bounded by (Γ). Nevertheless, there is a good sense in which M still has a minimum area property, namely: if M' denote any portion of M, which with its boundaries Γ' is completely interior to M, then the area of M' is finite and a minimum with respect to all surfaces of the same topological type bounded by Γ'.[62]

Choosing, for illustration, the circular representation of R^*, let C' be a system of k circles concentric with the circles C which correspond to the contours Γ, and slightly larger or smaller than C so as to be in the interior of R^*. Then C' bounds a portion[63] $R^{*\prime}$ of R^*, and this corresponds on M to a portion M', completely interior, including its boundaries Γ' (which are k analytic curves). Also, when m is sufficiently large, C' is evidently interior to the circular region R_m^*, and therefore determines on the minimal surface M_m bounded by Γ_m a completely interior portion M_m', inclusive of its boundaries Γ_m'.

First, we easily see that the area of M' is finite. For in the present case of a conformal representation of M on R^*, with $E = G$, $F = 0$, the area is expressed by the Dirichlet integral

$$\mathfrak{A}(M') = \int\int \frac{1}{2}(E+G)\,du\,dv = \int\int \frac{1}{2}\sum_{i=1}^{n}\left[\left(\frac{\partial H_i^*}{\partial u}\right)^2 + \left(\frac{\partial H_i^*}{\partial v}\right)^2\right] du\,dv. \tag{17.28}$$

As is well-known, the derivatives of the harmonic functions H_i^* remain continuous and bounded in the closed region $R^{*\prime}$; therefore $\mathfrak{A}(M')$ is finite.

The same applies to the total length $l(\Gamma')$ of the k boundaries Γ'. For

$$\begin{aligned} l(\Gamma') &= \int_{C'} \sqrt{(E\,du^2 + 2F\,du\,dv + G\,dv^2)} \\ &= \int_{C'} \sqrt{\sum_{i=1}^{n}\left[\frac{\partial H_i^*}{\partial u}\,du + \frac{\partial H_i^*}{\partial v}\,dv\right]^2}; \end{aligned} \tag{17.29}$$

therefore $l(\Gamma')$ is finite.

Further, we have

$$\lim_{m\to\infty} \mathfrak{A}(M_m') = \mathfrak{A}(M'), \tag{17.30}$$

and

$$\lim_{m\to\infty} l(\Gamma_m') = l(\Gamma'). \tag{17.31}$$

This is because the approach of the partial derivatives of H_m^* to those of H^* is uniform in the closed sub-region $R^{*\prime}$, a consequence of the boundedness of $g_m^*(z)$ in its approach to $g^*(z)$.

---

[62] This was proved for $k = 1$, $r = 0$ in the author's paper [8a]. The following is along the lines of the exposition by T. Radó [15], p. 94.

[63] This is the one instance in this paper where the accent does not denote the semi Riemann surface, but only a certain part of it.

Since the approach of H_m^* to H^* is uniform on C', the points of Γ_m' tend uniformly to the corresponding points of Γ'.

Let us then form the set of k ring-shaped surfaces ρ_m by joining corresponding points of Γ_m', Γ'. If δ_m denote the greatest distance between these corresponding points, then an upper bound for the area of ρ_m is evidently expressed by

$$\mathfrak{A}(\rho_m) \leqq \tfrac{1}{2}\delta_m [l(\Gamma_m') + l(\Gamma')]. \tag{17.32}$$

Since $\delta_m \to 0$, $l(\Gamma_m') \to l(\Gamma')$ and $l(\Gamma')$ is finite, we have

$$\lim_{m\to\infty} \mathfrak{A}(\rho_m) = 0. \tag{17.33}$$

Suppose now that it were possible to have a surface S' of the topological type of M' and bounded by Γ', such that

$$\mathfrak{A}(S') < \mathfrak{A}(M'). \tag{17.34}$$

Then let us form the surface

$$\Sigma_m = (M_m - M_m') + \rho_m + S'; \tag{17.35}$$

evidently this is bounded by Γ_m and is of topological type r. We have

$$\mathfrak{A}(\Sigma_m) = \mathfrak{A}(M_m) - \mathfrak{A}(M_m') + \mathfrak{A}(\rho_m) + \mathfrak{A}(S'), \tag{17.36}$$

or

$$\mathfrak{A}(\Sigma_m) - \mathfrak{A}(M_m) = \mathfrak{A}(S') - \mathfrak{A}(M_m') + \mathfrak{A}(\rho_m). \tag{17.37}$$

Hence by (17.30, 33, 34),

$$\lim_{m\to\infty} \{\mathfrak{A}(\Sigma_m) - \mathfrak{A}(M_m)\} = \mathfrak{A}(S') - \mathfrak{A}(M') < 0. \tag{17.38}$$

Accordingly, for sufficiently large values of m,

$$\mathfrak{A}(\Sigma_m) < \mathfrak{A}(M_m). \tag{17.39}$$

But this contradicts the minimum area property of M_m, as expressed by Theorem I. Therefore the hypothesis (17.34) is untenable; we have, on the contrary,

$$\mathfrak{A}(M') \leqq \mathfrak{A}(S) \tag{17.40}$$

for every surface S of topological type r bounded by Γ'.

Evidently the same minimum area property applies to any completely interior portion M'' of M, of whatever topological structure. For M'' can obviously be included in a region such as M', and if the area of M'' could be lessened with preservation of its boundaries and topological form, the same operation would lessen that of M' without disturbing its boundaries or topological form.

With this remark, the proof of Theorem III is completed.

5. Examples of contours which bound no finite area

Probably the simplest is the one defined in spherical polar coördinates by the following equations[64]

$$\rho = \cos\varphi, \qquad \theta = \tan^5\varphi, \tag{17.41}$$

where ρ = radius, φ = latitude, θ = longitude. This has the form of a spiral with asymptotic point at the origin, which is approached from above and below. Fig. 24 shows the orthogonal projection on the plane $z = 0$, where the arc drawn full corresponds to $z > 0$, and the dotted arc to $z < 0$. The origin is the only

FIG. 24

point of non-analyticity, and the infinite area of the minimal surface M is concentrated in its vicinity. The infinite area property is due essentially to the order of slowness of contraction of the coils of the spiral, as expressed by the divergence of the integral

$$\frac{1}{2}\int_{-\infty}^{+\infty} r^2\, d\theta = +\infty, \tag{17.42}$$

representing the area enclosed by the projection of the spiral on the xy-plane. This area is to be considered as a Riemann surface over the xy-plane, swept out by the radius vector r. We have $r = \rho\cos\varphi = \cos^2\varphi$; therefore as $\theta \to \pm\infty$, $r \sim \theta^{-4/5}$ (asymptotically), which implies the condition (17.42). It is evident that by means of this condition, examples of non-finite-area-bounding contours can be multiplied indefinitely.

Since the area of any completely interior portion M' of M is finite, the infinite part of the area of M lies always adjacent to its boundary, due, as we may say intuitively, to an increasing curliness of the surface as we approach the boundary. In the preceding example, the infinite part of the area is even concentrated near a single point of the contour.

The author has also given an example where infinite area occurs in the vicinity of every boundary point.[65] This is the contour defined by the equations

$$x = \sum_{m=1}^{\infty} \frac{1}{2^m}\cos 4^m\varphi, \qquad x = \sum_{m=1}^{\infty} \frac{1}{2^m}\sin 4^m\varphi, \qquad z = \sin\varphi. \tag{17.43}$$

[64] Douglas [9].

[65] We have also constructed a more geometric type of example, which we intend to publish soon. (Added in proof: since appeared in Proc. Nat. Acad. Sci., v. 24, no. 10, Oct., 1938, pp. 490–495.)

The xy-projection is the limit of an epicyclic curve with an increasingly dense distribution of loops. Combined with the z-equation, these loops produce sufficiently slowly contracting spirals, like that of the previous example, which are distributed in an everywhere dense manner along the entire contour. Roughly speaking, any singularity presented by this contour will be repeated on an everywhere dense set of points, since the transformation of parameter

$$\varphi \text{ into } \varphi + \frac{2p\pi}{4^q} \qquad (p, q \text{ any positive integers})$$

has no effect on the terms $m \geqq q$, and therefore changes the series for x and y only by the addition of a finite trigonometric sum. The fractions $p/4^q$ form an everywhere dense set.

§18. Conformal Mapping of Plane Regions

1. Conformal mapping as contained in the Plateau problem

From the start, it has been a feature of the author's work on the Plateau problem that the conformal mapping of Jordan regions in the plane is included automatically as the case $n = 2$, as well as the one-one continuous relation between boundaries induced by the conformal map.

This continues to apply for a general value of the connectivity k of the region, so that the present work includes th. classic theorem of Schottky [24] concerning the conformal mapping on a Riemann surface of a plane region J_k bounded by any k Jordan curves (Γ).[66]

In fact, if $n = 2$, the condition $\sum_{i=1}^{n} F_i'^2(w) = 0$ becomes

$$F_1'(w) = \pm i F_2'(w). \tag{18.1}$$

Let

$$x_1 = P_1(u, v) = \Re F_1(w), \qquad x_2 = P_2(u, v) = \Re F_2(w); \tag{18.2}$$

then

$$F_1'(w) = \frac{\partial P_1}{\partial u} - i\frac{\partial P_1}{\partial v}, \qquad F_2'(w) = \frac{\partial P_2}{\partial u} - i\frac{\partial P_2}{\partial v}, \tag{18.3}$$

so that, using the plus sign[67] in (18.1), we have

$$\frac{\partial P_1}{\partial u} = \frac{\partial P_2}{\partial v}, \qquad \frac{\partial P_1}{\partial v} = -\frac{\partial P_2}{\partial u}. \tag{18.4}$$

These are the Cauchy-Riemann equations. Therefore $x_1 = ix_2 + F(w)$, an analytic function of w, regular in the interior of the semi Riemann surface R', remaining continuous on its boundaries C and converting these topologically into the contours Γ. It follows by a theorem of Darboux,[68] that $F(w)$ transforms

[66] We suppose these sensed so that J_k is on the left.

[67] The minus sign leads to an inversely conformal transformation.

[68] See Osgood, *Funktionentheorie*, 1923, p. 377. This Darboux theorem is essentially a corollary of the Cauchy indicatrix theorem.

the region R' topologically and conformally, without singular points ($F'(w) \neq 0$), into the given region J_k.

In order to be able to apply to $n = 2$ the preceding results established for any value of n, we must show that the sufficient condition $d(\Gamma) < d'(\Gamma)$ is verified.[69] It will be evident that we need only provide against a separation of (Γ) into a set of 1 and a set of $k - 1$, or, typically, to prove

$$d(\Gamma_1, \cdots, \Gamma_k) < d(\Gamma_1, \cdots, \Gamma_{k-1}) + d(\Gamma_k). \tag{18.5}$$

If, without use of any extraneous theory of conformal mapping, we could prove the equality of d(Γ) and the minimum area $a(\Gamma)$, then this relation would be obvious; in fact, we even have

$$a(\Gamma_1, \cdots, \Gamma_k) = a(\Gamma_1, \cdots, \Gamma_{k-1}) - a(\Gamma_k), \tag{18.6}$$

or yet more strongly, with an additional term: minus the content of Γ_k, in case the curve Γ_k has positive content.[70]

But in order to show that $d(\Gamma) = a(\Gamma)$, we had to use the conformal mapping of polyhedral surfaces. Therefore the point is to establish the sufficient condition (18.5) directly on the basis of the definition of $d(\Gamma)$ as minimum of the functional $A(g, R)$.

2. Partitions $k = k' + k''$ and $k = 1 + (k - 1)$

We have stated it to be evident that the k conditions of the type (18.5) are sufficient to imply for *every* separation

$$(\Gamma) = (\Gamma') + (\Gamma''), \qquad k = k' + k'' \tag{18.7}$$

that

$$d(\Gamma) < d(\Gamma') + d(\Gamma''), \tag{18.8}$$

Essentially, this is because whenever each component set contains more than a single contour, the set, say (Γ'), which does not contain the external boundary of J_k consists of $k' > 1$ mutually external plane curves. The minimal surface M determined by these obviously reduces to the sum of the plane discs bounded by the individual contours. This gives

$$d(\Gamma') = d(\Gamma'_1) + d(\Gamma'_2) + \cdots + d(\Gamma'_{k'}). \tag{18.9}$$

Consequently, since $\leqq$ always holds in relations like (18.8), we have

$$\begin{aligned} d(\Gamma') + d(\Gamma'') &= d(\Gamma'_1) + d(\Gamma'_2) + \cdots + d(\Gamma'_{k'}) + d(\Gamma'') \\ &\geqq d(\Gamma'_1) + d(\Gamma'_2, \cdots, \Gamma'_k; \Gamma) > d(\Gamma), \end{aligned} \tag{18.10}$$

the last by the assumed conditions of type (18.5).

---

[69] Since throughout this section $r = 0$, it is omitted.

[70] For examples of Jordan curves of positive content, see Osgood, Trans. Amer. Math. Soc., v. 4 (1903), pp. 107–112; also K. Knopp, Archiv. Math. und Physik., (1918).

This reasoning provides a substantial heuristic basis, but we wish to emphasize that the proof really depends only on the functional $A(g, R)$ and not on any interpretation of $d(\Gamma)$ as minimum area.

To this end, let first Γ_1', Γ_2' denote any two mutually external plane contours, similarly sensed. Then we must have

$$d(\Gamma_1', \Gamma_2') = d(\Gamma_1') + d(\Gamma_2'). \tag{18.11}$$

For, by §14.2, the only alternative is

$$d(\Gamma_1', \Gamma_2') < d(\Gamma_1') + d(\Gamma_2'). \tag{18.12}$$

This is the unique sufficient condition occuring in the two-contour case. It would imply the existence of a regular analytic function $F(w)$ in a circular ring C_1C_2, where the values of $F(w)$ cover exactly the sum of the interiors of Γ_1' and Γ_2'. But this is impossible, for the one domain is connected, the other disconnected.

To deal with the case $k > 2$, we use recurrent reasoning. Suppose that the relation (18.9) has already been established for any set of mutually external, similarly sensed contours (Γ') less than k' in number. Let

$$(\Gamma') = (\Gamma')_1 + (\Gamma')_2$$

denote any separation of (Γ'). Then for at least one such separation, we must have

$$d(\Gamma') = d(\Gamma')_1 + d(\Gamma')_2 . \tag{18.13}$$

For if we always had the relation $<$ between the two members, this would express the condition sufficient for the existence of a topological and conformal map of some connected circular domain or Riemann surface on the disconnected domain consisting of the interiors of $\Gamma_1', \cdots, \Gamma_{k'}'$, just as in the case $k' = 2$, and just as absurd. Now, each set $(\Gamma')_1$, $(\Gamma')_2$ in less than k' in number, and so by hypothesis the relation (18.9) applies to $(\Gamma')_1$, $(\Gamma')_2$; consequently, by (18.13), to (Γ').

Thus the proof is completed that the particular relations of type (18.5) imply the general relations (18.8).

Proof of the sufficient condition (18.5)

3. Asymptotic formula for $A(g, g; q)$

In "Two Contours" we established the sufficient condition

$$d(\Gamma_1, \Gamma_2) < d(\Gamma_1) + d(\Gamma_2) \tag{18.14}$$

by means of an asymptotic formula[71] expressing the functional $A(g_1, g_2; q)$ for

[71] [3], p. 339, formula (5.6).

small values of q, where $q < 1$ is the radius of C_2, unity the radius of the concentric circle C_1. Using this formula, we proved that if

$$A(g_1^*) = \min A(g_1) = d(\Gamma_1), \qquad A(g_2^*) = \min A(g_2) = d(\Gamma_2), \tag{18.15}$$

and if g_1^*, g_2^* are properly normalized by adjustment of the three parameters involved in each, then *for all sufficiently small values of* q,

$$A(g_1^*, g_2^*; q) < A(g_1^*) + A(g_2^*). \tag{18.16}$$

Since, by definition,

$$d(\Gamma_1, \Gamma_2) \leqq A(g_1^*, g_2^*; q), \tag{18.17}$$

the sufficient condition (18.14) follows directly from (18.15, 16).

Explicitly, let Γ_1 surround Γ_2, let P denote any interior point of Γ_2—therefore of Γ_1, and L any directed element through P. By the theory of the one-contour case, g_1^* and g_2^* determine conformal maps of the unit circle on the regions bounded by Γ_1 and Γ_2 respectively; let $F_1(w)$, $F_2(w)$ denote the analytic functions defining these maps. We suppose them to be so normalized that each converts the origin O into P, and the direction at O of the positive real axis into L.

On account of the interior regularity of the conformal maps, we have

$$\mu_1 = |F_1'(0)| > 0, \qquad \mu_2 = |F_2'(0)| > 0. \tag{18.18}$$

Then the stated asymptotic formula is

$$A(g_1^*, g_2^*; q) = A(g_1^*) + A(g_2^*) - 4\pi\mu_1\mu_2 q + \cdots, \tag{18.19}$$

where the remaining terms form a convergent series in higher powers of q.

Evidently, for sufficiently small values of q, the term in the first power of q is dominant. Since its coefficient is, by (18.18), strictly negative, the result (18.16) follows directly.

4. Reduction of the general case to $k = 2$

The proof of the sufficient condition (18.5) for a general value of k, can be made to depend on the case $k = 2$ by the following procedure, based on recurrent reasoning.

Suppose the conformal mapping problem already solved for $k - 1$, and thereby let the Jordan region J_{k-1} bounded by $\Gamma_1, \cdots, \Gamma_{k-1}$ be conformally represented on the circular region R_{k-1}^* bounded by $C_1, \cdots, C_{k-1}$. The figure illustrates this, as well as the following constructions, for $k = 4$.

Then denoting by $g_1^*, \cdots, g_{k-1}^*$ the one-one continuous representation of $\Gamma_1, \cdots, \Gamma_{k-1}$ on $C_1, \cdots, C_{k-1}$ corresponding to the conformal map, we have obviously,

$$A(g_1^*, \cdots, g_{k-1}^*; R_{k-1}^*) = d(\Gamma_1, \cdots, \Gamma_{k-1}). \tag{18.20}$$

Now adjoin Γ_k to the region J_{k-1}, and choose arbitrarily a point P interior to Γ_k together with L any fixed direction through P. By the conformal map of J_{k-1} on R^*_{k-1}, these will correspond respectively to a point p and a direction l. About p as center construct a circle C_k with a radius q that will tend to zero.

Let the region interior to Γ_k be mapped conformally on the circular disc C_k so that the directed lineal element P, L corresponds to p, l, which is possible because of the three parameters involved. This mapping will be accomplished by means of a minimizing parametric representation $g^\dagger_k$ of Γ_k on C_k :

$$A(g^\dagger_k) = d(\Gamma_k). \tag{18.21}$$

Denote by $R_k(q)$ the circular region obtained by removing the circular disc C_k from R^*_{k-1}. Then we shall prove the following, which generalizes (18.16): *for all sufficiently small values of* q,

$$A(g^*_1, \cdots, g^*_{k-1}, g^\dagger_k; R_k(q)) < A(g^*_1, \cdots, g^*_{k-1}; R^*_{k-1}) + A(g^\dagger_k). \tag{18.22}$$

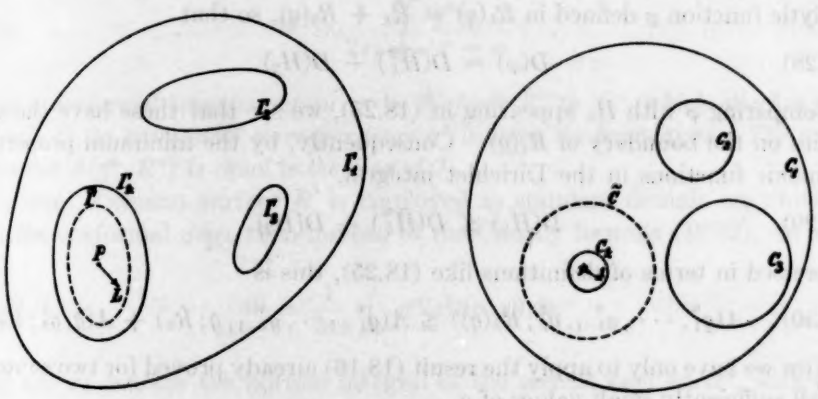


FIG. 25

When this is done, the desired sufficient condition (18.5) is an immediate consequence, by (18.20, 21) and the definition of $d(\Gamma_1, \cdots, \Gamma_k)$.

5. Proof of (18.22)

Let $\tilde{C}$ denote a circle about p as center, and $\tilde{\Gamma}$ its image in the conformal map of J_{k-1} on R^*_{k-1}. We suppose the radius of $\tilde{C}$ to be fixed, but small enough so that $\tilde{\Gamma}$ is completely interior to Γ_k. $\tilde{\Gamma}$ appears in a definite parametric representation $\tilde{g}$ on $\tilde{C}$.

Denote by $\tilde{R}_1$ the circular disc $\tilde{C}$, and by $R_2(q)$ the ring $\tilde{C}C_k$. Let $\tilde{R}_k = R^*_{k-1} - \tilde{R}_1$. We have then

$$R^*_{k-1} = \tilde{R}_k + \tilde{R}_1, \tag{18.23}$$

$$R_k(q) = \tilde{R}_k + R_2(q). \tag{18.24}$$

Construct now the harmonic vector function H_k in $R_k(q)$ with the boundary values $g_1^*, \cdots, g_{k-1}^*$ on $C_1, \cdots, C_{k-1}$ and $g_k^\dagger$ on C_k. Then by definition,

$$A(g_1^*, \cdots, g_{k-1}^*, g_k^\dagger; R_k(q)) = D(H_k). \tag{18.25}$$

Let H_{k-1}^* denote the harmonic vector function in R_{k-1}^* which defines the conformal map on J_{k-1}, so that H_{k-1}^* has the boundary values $g_1^*, \cdots, g_{k-1}^*$ on $C_1, \cdots, C_{k-1}$. With regard to (18.23), denote by $\tilde{H}_k^*$ the part of H_{k-1}^* belonging to $\tilde{R}_k$, and by $\tilde{H}_1^*$ the part belonging to $\tilde{R}_1$. Then

$$D(H_{k-1}^*) = D(\tilde{H}_k^*) + D(\tilde{H}_1^*); \tag{18.26}$$

so that, by the definitions analogous to (18.25),

$$A(g_1^*, \cdots, g_{k-1}^*; R_{k-1}^*) = A(g_1^*, \cdots, g_{k-1}^*, \tilde{g}_k; \tilde{R}_k) + A(\tilde{g}). \tag{18.27}$$

Construct next the harmonic function H_2 in the ring $R_2(q)$ with the boundary values $\tilde{g}$ on $\tilde{C}$, $g_k^\dagger$ on C_k. By the previous definitions, $\tilde{g}$ is the value in $\tilde{C}$ of $\tilde{H}_k^*$. Accordingly, $\tilde{H}_k^*$ attaches continuously to H_2, and forms with it a piece-wise analytic function φ defined in $R_k(q) = \tilde{R}_k + R_2(q)$, so that

$$D(\varphi) = D(\tilde{H}_k^*) + D(H_2). \tag{18.28}$$

Comparing φ with H_k appearing in (18.25), we see that these have the same values on the boundary of $R_k(q)$. Consequently, by the minimum property of harmonic functions in the Dirichlet integral,

$$D(H_k) \leqq D(\tilde{H}_k^*) + D(H_2). \tag{18.29}$$

Expressed in terms of definitions like (18.25), this is

$$A(g_1^*, \cdots, g_{k-1}^*, g_k^\dagger; R_k(q)) \leqq A(g_1^*, \cdots, g_{k-1}^*, \tilde{g}; \tilde{R}_k) + A(\tilde{g}, g_k^\dagger; R_2(q)). \tag{18.30}$$

Now we have only to apply the result (18.16) already proved for two contours: for all sufficiently small values of q,

$$A(\tilde{g}, g_k^\dagger; R_2(q)) < A(\tilde{g}) + A(g_k^\dagger). \tag{18.31}$$

The proper normalization of $\tilde{g}$, $g_k^\dagger$ necessary for the application of (18.16) is secured by our arrangement that the corresponding conformal maps: the disc $\tilde{\Gamma}$ on $\tilde{C}$, the disc Γ_k on C_k, both convert the point and direction P, L into the point and direction p, l.

Finally, we apply (18.31) in (18.30) and then combine the first two terms on the right according to (18.27). The result is the desired inequality (18.22), and with it, as stated, the sufficient condition (18.5).

6. Conformal mapping theorem

We have thus proved by means of the functional $A(g, R)$ the following theorem.[72]

[72] [1], Theorem III.

THEOREM IV. *Any given plane region J_k, bounded by any finite number k of Jordan curves Γ, can be mapped conformally on the semi Riemann surface R' of an algebraic curve of genus $k-1$ and with k real branches, or on the equivalent domain bounded by k circles. The conformal map attaches continuously to a topological correspondence between the boundaries.*

In order that two Jordan regions J_k, $\tilde{J}_k$ may be mapped conformally on each other, it is necessary and sufficient that the corresponding algebraic curves R, $\tilde{R}$ be birationally equivalent.

If the (x_1, x_2) plane be regarded as a complex plane, Z or $W = x_1 + ix_2$, and the (u, v) domain as locus of the complex variable z or $w = u + iv$, then we may summarize as follows.

Let $Z = g(z)$ denote an arbitrary parametric representation of the k contours Γ bounding a given Jordan region J_k, on the k circles C bounding a variable circular domain R. A particular representation g^*, R^* exists which minimizes the functional $A(g, R)$. Then the integral formula of Cauchy:

$$W = \frac{1}{2\pi i}\int_{C^*} \frac{g^*(z)\,dz}{z-w}, \tag{18.32}$$

defines a conformal transformation $w \to W$ from R^* to J_k, which attaches continuously to the topological correspondence g^* between the boundaries. The minimum value $A(g^*, R^*)$ is equal to the area of J_k.

If a semi Riemann surface R' is employed as standard domain on which to make the conformal map, then instead of the Cauchy formula (18.32), we have

$$W = \frac{1}{2\pi i}\int_{C^*} g^*(z)t(z, w)\,dz, \tag{18.33}$$

where $t(z, w)$ denotes the normal integral of the second kind on R^*, with pole at w, and having zero periods with respect to the real branches C.

The preceding theory applies also to the case where the Jordan curves Γ have points in common, provided that the region J_k remains connected. (Cf. §3.3, and [8].)

Theorem IV goes back to Schottky, whose proof however did not include the topological correspondence between boundaries. The present mode of proof, in which, by means of the functional $A(g, R)$, we find first the appropriate standard region R^* and the correspondence g^* between boundaries, was given by the author for $k = 1$ in [5], for $k = 2$ in [3], and for a general value of k in [1], [2].

The interior conformal map is deduced immediately from g^*, R^*, as described. Previously, the interior and boundary maps required separate treatment. For $k = 1$, the existence of the interior map was established essentially by Riemann [25]. The nature of the boundary correspondence was proved much later by Osgood [22] and by Carathéodory [23], independently.

§19. Alternative Method

1

If we relinquish the object of including in our method the classic theorems on the conformal mapping of plane regions, then we can simplify our treatment considerably by using *from the start* the theorem that it is possible to map any polyhedral surface conformally on a semi Riemann surface of like topological structure, the map remaining one-one continuous on the boundaries. In presenting this alternative method, we shall follow—with the appropriate generalizations—the lines of our paper [7], which treats the case $k = 1$, $r = 0$.[73]

2. Proof

Hypotheses:

(α) $$a(\Gamma, r) < + \infty,$$

(β) $$a(\Gamma, r) < a(\Gamma, r').$$

The proof to be given is logically self-contained. We collect from the preceding developments the following list of quite simple facts, which constitute the complete materials out of which our proof will be constructed. They are arranged in the order of use in the proof, and will be referred to by the number here given.

(1°) The Lebesgue definition of area (16.1):

$$\mathfrak{A}(S) = \min \lim_{j \to \infty} \mathfrak{S}(\Pi_j),$$

where the polyhedral surface Π_j tends to S in the Fréchet sense.

(2°) $d(\Gamma, r) \geqq a(\Gamma, r)$, $d(\Gamma, r') \geqq a(\Gamma, r')$. (5.25, 25′).

(3°) The polyhedral mapping theorem just stated.

(4°) Theorem (a), §5. $D(H) \geqq I(H)$. If $D(H) = I(H)$, then $E \equiv G$, $F \equiv 0$, and conversely.

(5°) The minimum property of the Dirichlet integral: $D(H) \leqq D(\mathfrak{x})$, if the harmonic vector H has the same boundary values as $\mathfrak{x}$.

(6°) The compactness and closure properties of the set $[(\Gamma, g, R)]$ for suitably varying as well as for fixed contours (Γ). §13.

(7°) $A(g, R)$ is lower semi-continuous. §14.

(8°) $d(\Gamma, r) = d(\Gamma, r, r'') \leqq A(g, R)$ for *all* representations (g, R), proper and improper. (14.8).

(9°) If g is discontinuously improper, then $A(g, R) = + \infty$. (13.1).

(10°) If two transition curves of R intersect, and the corresponding contours do not, then $A(g, R) = + \infty$. (13.1).

(11°) If g is degenerate, then $A(g, R) \geqq d(\Gamma, r')$. (14.32).

---

[73] See also T. Radó [14], [15] for the case $k = 1$, $r = 0$. The conformal mapping of polyhedra, as method, was introduced into the Plateau problem by Radó.

(12°) The exact definition of a surface S_r bounded by (Γ). §4.2.

(13°) $\mathfrak{A}(H) \leqq I(H)$. (16.11).

(14°) $a(\Gamma, r) = a(\Gamma, r, r'') \leqq \mathfrak{A}(S)$ for all surfaces S, whether of proper or reduced type (r, r''), bounded by (Γ). (5.5).

(15°) If the harmonic surface H determined by the representation (g, R) is in conformal representation on R, i.e. $E = G$, $F = 0$, then g cannot be continuously improper. §15.4.

3

We begin by defining with Lebesgue

$$m(\Gamma, r) = \min \lim_{j\to\infty} \mathfrak{S}(\Pi_j), \tag{19.1}$$

where the minimum is with respect to all sequences of polyhedral surfaces Π_j of proper topological type r, whose polygonal boundaries $(P)_j$ tend to (Γ) in the Fréchet sense (16.4). Nothing is stipulated as to the behavior of the polyhedral surfaces themselves.

Then it is evident by (1°) that for every surface S_r of topological type r, bounded by (Γ), we have

$$m(\Gamma, r) \leqq \mathfrak{A}(S_r), \tag{19.2}$$

since if $\Pi_j \to S_r$, then $(P)_j \to (\Gamma)$. Therefore

$$m(\Gamma, r) \leqq a(\Gamma, r) \leqq d(\Gamma, r), \tag{19.3}$$

by (2°).

It follows easily from the definition (19.1) that we can find a particular sequence Π_j where $(P)_j$ tends to (Γ) and the area $\mathfrak{S}(\Pi_j)$ tends exactly to the minimum possible value:

$$\lim_{j\to\infty} \mathfrak{S}(\Pi_j) = m(\Gamma, r). \tag{19.4}$$

For, given $\epsilon > 0$, we can find by (19.1) a particular sequence such that

$$m(\Gamma, r) \leqq \underline{\lim}\ \mathfrak{S}(\Pi_j) \leqq m(\Gamma, r) + \epsilon;$$

and therefore can select from this a particular element Π_ϵ with boundary $(P)_\epsilon$ such that

$$\text{dist}\,(P_\epsilon, \Gamma) < \epsilon, \qquad m(\Gamma, r) - \epsilon \leqq \mathfrak{S}(\Pi_\epsilon) \leqq m(\Gamma, r) + 2\epsilon.$$

By allowing ϵ to run through a sequence $\epsilon_j \to 0$, we arrange (19.4).

Let Π_j now be represented conformally, according to (3°) on a semi Riemann surface R_j' of topological type r. This constitutes a certain parametric representation of Π_j, defined by a vector function on R_j', which vector function we denote again by Π_j. The representation being conformal, we have $E \equiv G$, $F \equiv 0$; therefore by (4°),

$$D(\Pi_j) = I(\Pi_j) = \mathfrak{S}(\Pi_j). \tag{19.5}$$

Let H_j denote the harmonic vector function on R_j', with the same boundary values g_j as Π_j, where g_j is then a certain parametric representation of $(P)_j$; then by (5°)

$$A(g_j, R_j) = D(H_j) \leqq D(\Pi_j) = \mathfrak{E}(\Pi_j). \tag{19.6}$$

Since $(P)_j \to (\Gamma)$, the sequence (P_j, g_j, R_j) is compact (6°), and we can therefore select a partial sequence—denote it without change of index—such that

$$(P_j, g_j, R_j) \to (\Gamma, g_0, R_0), \tag{19.7}$$

where (g_0, R_0) is a certain representation of the contours (Γ), proper or improper.

By (7°), we have therefore, with use of (19.6, 4, 3),

$$\begin{aligned} A(g_0, R_0) &\leqq \lim_{j\to\infty} A(g_j, R_j) \leqq \lim_{j\to\infty} \mathfrak{E}(\Pi_j) \\ &= m(\Gamma, r) \leqq a(\Gamma, r) \leqq d(\Gamma, r). \end{aligned} \tag{19.8}$$

By (8°), regardless of whether (g_0, R_0) is proper or improper, we have also the reverse inequality:

$$A(g_0, R_0) \geqq d(\Gamma, r).$$

It follows that

$$A(g_0, R_0) = m(\Gamma, r) = a(\Gamma, r) = d(\Gamma, r). \tag{19.9}$$

Since by hypothesis (α), $a(\Gamma, r) < +\infty$, we can therefore, by reference to (9°), exclude the possibility that (g_0, R_0) is discontinuously improper. Also, by (10°), the possibility that two transition curves of R_0 intersect is excluded. We are assuming that no two contours Γ intersect.

If g_0 is degenerate, then by (11°), (2°) and hypothesis (β),

$$A(g_0, R_0) \geqq d(\Gamma, r') \geqq a(\Gamma, r') > a(\Gamma, r), \tag{19.10}$$

which contradicts (19.9). Therefore this possibility is also excluded.

There remain to be excluded only the two possibilities that R_0 be reduced, or that g_0 be continuously improper.

Without prejudice of these possibilities, certainly we can construct a harmonic vector H_0 on the semi-surface R_0' with the boundary values g_0. Then, according to (12°), H_0 represents a surface of topological type r or r'' bounded by (Γ). Therefore, by (13°), (4°) and (19.9),

$$\mathfrak{A}(H_0) \leqq I(H_0) \leqq D(H_0) = A(g_0, R_0) = a(\Gamma, r). \tag{19.11}$$

But, on the other hand, we have by (14°) the reverse inequality

$$\mathfrak{A}(H_0) \geqq a(\Gamma, r), \tag{19.12}$$

regardless of the, as yet undecided, proper or reduced character of H_0. Hence

$$\mathfrak{A}(H_0) = I(H_0) = D(H_0) = a(\Gamma, r). \tag{19.13}$$

Therefore H_0, i.e. R_0, *is not reduced*, for if it were, then with hypothesis (β) we would have

$$\mathfrak{A}(H_0) \geqq a(\Gamma, r') > a(\Gamma, r), \tag{19.14}$$

contradicting (19.13).

Thus H_0 is a harmonic surface of proper topological type r, bounded by (Γ).

By (19.13), $D(H_0) = I(H_0)$; therefore applying (4°), it follows that the representation of H_0 on R_0 is conformal:

$$E \equiv G, \qquad F \equiv 0. \tag{19.15}$$

But this is exactly the condition which makes the harmonic surface H_0 a minimal surface.

That H_0 has minimum area is already expressed in (19.13).

Finally, we see by (15°) that the correspondence between the boundaries (C_0) of R_0' and (Γ) of H_0 is one-one.

§20. Concluding Remarks

The developments of this paper can be extended to calculus of variations problems of quite general form, where we seek an extremal surface of prescribed topological type with any finite number of given boundaries. Sufficient conditions like $d(\Gamma, r) < d(\Gamma, r')$ prevent the degeneration of the extremal surface to a lower topological type or separation into a number of disconnected parts. Here $d(\Gamma, r)$ denotes, of course, the lower bound for surfaces with given boundaries (Γ) and assigned topological type r, of whatever integral is to be minimized in the particular problem. r' indicates (primary) reduction or separation of the surface. One-sided surfaces are included quite as well as two-sided.

§21. Bibliography

1. J. Douglas, *Some new results in the problem of Plateau*, Journal of Math. and Physics, v. 15 (1936), pp. 55–64.
2. J. Douglas, *Minimal surfaces of general topological structure with any finite number of assigned boundaries*, Ibid., v. 15 (1936), pp. 106–123.
3. J. Douglas, *The problem of Plateau for two contours*, Ibid., v. 10 (1931), pp. 310–359.
4. J. Douglas, *One-sided minimal surfaces with a given boundary*, Trans. Amer. Math. Soc., v. 34 (1932), pp. 731–756.
5. J. Douglas, *Solution of the problem of Plateau*, Ibid., v. 33 (1931), pp. 263–321.
6. J. Douglas, *The problem of Plateau*, Bulletin Amer. Math. Soc., v. 39 (1933), pp. 227–251.
7. J. Douglas, *The mapping theorem of Koebe and the problem of Plateau*, Journal of Math. and Physics, v. 10 (1931), pp. 106–130.
8. J. Douglas, *Crescent-shaped minimal surfaces*, Proc. Nat. Acad. Sci., U. S. A., v. 19 (1933), pp. 192–199.

8a. J. Douglas, *The least area property of the minimal surface determined by an arbitrary Jordan contour*, Ibid., v. 17 (1931), pp. 211–216.

9. J. Douglas, *An analytic closed space curve that bounds no orientable surface of finite area*, Ibid., v. 19 (1933), pp. 448–451.

10. J. Douglas, *A Jordan curve no arc of which can form part of a contour which bounds a finite area*, Annals of Math., v. 35 (1934), pp. 100–104.
11. R. Courant, *On the problem of Plateau*, Proc. Nat. Acad. Sci., U. S. A., v. 22 (1936), pp. 367–372.
12. R. Courant, *Plateau's problem and Dirichlet's principle*, Annals of Math., v. 38 (1937), pp. 679–724.
13. L. Tonelli, *Sul problema di Plateau*, Rendiconti Lincei, v. 24 (1936), pp. 333–339, 393–398.
14. T. Radó, *The problem of least area and the problem of Plateau*, Math. Zeitschrift, v. 32 (1930), pp. 763–796.
15. T. Radó, *On the Problem of Plateau*, Ergebnisse der Mathematik, 1933.
16. F. Klein, *Über Riemanns Theorie der algebraischen Funktionen und ihrer Integrale*, 1882.
17. H. F. Baker, *Abel's Theorem and the Allied Theory including the Theory of Theta Functions*, 1897.
18. E. Picard, *Traité d'Analyse*, v. 2, 1925, chaps. 14–16.
19. Severi-Löffler, *Vorlesungen über Algebraische Geometrie*, 1921.
20. B. v. Kerékjártó, *Vorlesungen über Topologie*, 1923.
21. Hilbert and Cohn-Vossen, *Anschauliche Geometrie*, 1932.
22. W. F. Osgood and E. H. Taylor, *Conformal transformations on the boundaries of their regions of definition*, Trans. Amer. Math. Soc., v. 14 (1913), pp. 277–298.
23. C. Carathéodory, *Über die gegenseitige Beziehung der Ränder bei der konformen Abbildung der Innern einer Jordanschen Kurve auf einen Kreis*, Math. Annalen, v. 73 (1913), pp. 305–320.
24. F. Schottky, *Über die konforme Abbildung mehrfach zusammenhängender ebener Flächen*, Crelle, v. 83 (1877), pp. 300–351.
25. B. Riemann, *Gesammelte Werke*, 1892, pp. 1–43.
26. M. Fréchet, *Sur quelques points du calcul fonctionnel*, Rendiconti Palermo, v. 22 (1906), pp. 1–74.
27. H. Lebesgue, *Intégrale, longueur, aire*, Annali di Matematica, v. 7 (1902), pp. 231–359.
28. W. Blaschke, *Vorlesungen über Differential-geometrie*, v. 1, 1921, chap. 6.
29. G. Darboux, *Leçons sur la Théorie Générale des Surfaces*, v. 1, 1887, pp. 267–506.
[74]30. J. Douglas, *Remarks on Riemann's doctoral dissertation*, Proc. Nat. Acad. Sci., v. 24 (1938), pp. 297–302.
31. J. Douglas, *Minimal surfaces of higher topological structure*, Ibid., v. 24 (1938), pp. 343–353. (Preliminary note to the present paper.)
32. J. Douglas, *Green's function and the problem of Plateau*, Ibid., v. 24 (1938), pp. 353–360.
33. J. Douglas, *The most general form of the problem of Plateau*, Ibid., v. 24 (1938), pp. 360–364.

[74] Nos. 30–33 published by the author since the submission to the editors of the manuscript of the present paper.